UNIVERSITE DE GENEVE
Département d'anthropologie et d'écologie

UNIVERSITE DE BERNE
Institut de zoologie

FACULTE DES SCIENCES

Professeur A. Langaney

Professeur L. Excoffier

Effets des expansions des populations humaines en Europe sur leur diversité génétique

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention biologique

par

Mathias Currat

de

Le Crêt (FR)

Thèse N° 3544

GENÈVE

Atelier de reproduction de l'Université de Genève
2004

La Faculté des sciences, sur le préavis de Messieurs L. EXCOFFIER, professeur et directeur de thèse (Université de Berne – Département de zoologie), A. LANGANEY, professeur ordinaire et co-directeur de thèse (Département d'anthropologie et écologie), L. CHIKHI, docteur (Université Paul Sabatier – Laboratoire Evolution et Diversité – Toulouse, France) et Madame A. SANCHEZ-MAZAS, professeur titulaire (Département d'anthropologie et écologie), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 10 août 2004

P. Spierer

Thèse - 3544 -

Le Doyen, Pierre SPIERER



à Christiane et René

Résumé

Ce travail de thèse décrit l'effet de l'expansion (spatiale et démographique) d'une population humaine sur sa diversité génétique, au moyen d'une approche par simulation. Le logiciel développé dans le cadre de cette étude est présenté de façon détaillée. Il est ensuite utilisé pour décrire la diversité génétique attendue dans une population qui est passée par une phase d'expansion, d'une part dans une aire inoccupée et d'autre part dans une aire déjà peuplée. Cette méthodologie est ensuite appliquée à deux cas particuliers d'expansion en Europe : celle des Hommes modernes entre 45'000 et 30'000 BP, et celle des populations néolithiques entre 10'000 et 5'000 BP. Ces recherches ont montré que la contribution des Néandertaliens au patrimoine génétique des humains modernes est vraisemblablement minimale, et que les gradients de fréquences alléliques observés dans les populations européennes ne sont pas une preuve de la migration des premiers agriculteurs néolithiques depuis le Proche-Orient.

Remerciements

La réalisation de cette thèse n'aurait pas été possible sans l'aide et les encouragements de nombreuses personnes. Leur soutien a pris des formes très diverses, et je tiens à leur témoigner ici ma plus profonde gratitude.

Laurent Excoffier a été à la fois l'instigateur et le superviseur des recherches présentées dans cette thèse. La qualité de mon travail doit beaucoup à son encadrement dynamique et motivant, à sa haute compétence et à sa rigueur scientifique. Sa soif de découverte et de changement m'a offert la possibilité de traverser la Sarine régulièrement et de savoir ce que "Töggelli" veut dire. Je le remercie également pour l'amitié qu'il m'a témoignée pendant ces années de fructueuse collaboration.

André Langaney m'a offert l'opportunité de poursuivre mon apprentissage dans son laboratoire à l'issue de mon diplôme et a ainsi rendu possible la réalisation de cette thèse. Son ouverture d'esprit et ses conseils inspirés m'ont été précieux pour franchir les différentes étapes qui ont constitué cette thèse et son regard critique m'a rendu attentif aux dérives du monde scientifique. Je le remercie également pour la grande liberté et la confiance qu'il m'a accordées dans la réalisation de mon travail.

Alicia Sanchez-Mazas m'a fait bénéficier de ses nombreuses connaissances dans différents domaines, notamment ceux des polymorphismes humains, de la préhistoire et de la linguistique. Elle m'a en particulier fait prendre concience de l'importance de la vulgarisation et de la diffusion au grand public des résultats de la recherche scientifique. Je la remercie également des nombreux conseils prodigués pendant mon apprentissage, ainsi que d'avoir accepté de faire partie de mon jury de thèse.

Lounès Chikhi a aimablement accepté d'être membre de mon jury de thèse. J'espère que l'avenir nous donnera l'opportunité d'une collaboration commune.

Nicolas Ray a réalisé son doctorat de façon parallèle au mien et notre collaboration a été quasiment quotidienne pendant près de quatre ans. Nous avons partagé non seulement de nombreux trajets en train et un repas médiéval mémorable, mais aussi les joies et les doutes quant au déroulement de nos travaux respectifs. Je tiens à le remercier pour sa gentillesse et son amitié.

Mes "officemate" successifs m'ont donné la possibilité, par leur enthousiasme et leur patience, de travailler dans un environnement agréable, motivant et dynamique. Je remercie particulièrement Sim Poloni pour sa gentillesse et sa disponibilité (je n'avais en effet pas épuisé mon stock de questions pendant mon diplôme) ainsi que Johan Renquin pour nos nombreuses discussions scientifiques et footballistiques. Je tiens à remercier également Lucia Simoni, Yann Beyer et Lele Reckeweg.

David Roessli a fait preuve d'une très grande disponibilité pour m'aider à résoudre les nombreux problèmes informatiques que j'ai rencontrés durant mon travail. Ses conseils avisés dans de nombreux autres domaines ont également été extrêmement bénéfiques à la finalisation de cette thèse.

Pierre Berthier n'a pas ménagé ses efforts pour mettre sur pied et entretenir de façon exemplaire le "cluster" informatique du CMPG. Sans son travail, les études présentées dans ce manuscrit n'auraient tout simplement pas pu voir le jour avant 2010 (au moins !).

Ma reconnaissance va également aux membres du LGB que j'ai eu la chance de côtoyer : Stéphane Bühler, Stefan Schneider, Isabelle Dupanloup, Ninian Hubert Van Blyenburgh, Patricia Dard, Jérôme Goudet, Alexandra Mossière, Rute Bucho, Barbara Arredi, Jose Manuel De Abreu Nunes.

Je tiens aussi à remercier les membres du CMPG, et plus particulièrement ceux avec qui j'ai eu le plaisir de collaborer dans le cadre du projet "Friction" : Samuel Neuenschwander, Grant Hamilton,

Guillaume Laval, Seraina Klopfstein et Daniel Wegmann, ainsi que Jean-Claude Nicod, Carlo Largiadèr, Gerald Heckel, Thomas Giger, Irene Keller, ainsi que les nombreux diplômants.

Les membres du Département d'Anthropologie et d'Ecologie de l'Université de Genève, à commencer par son ancien directeur Alain Gallay, mais également Marie Besse, Laurence-Isaline Stahl-Gretsch, Pierre-Yves Nicod, Jocelyne Desideri, Geneviève Perréard et Karoline Mazurié de Keroualin m'ont fait bénéficier de leurs compétences en archéologie et en anthropologie.

Le personnel technique et administratif du département n'a pas ménagé ses efforts pour que les miens puissent être focalisés sur la recherche. Je tiens à remercier particulièrement Jean-Gabriel Elia, Marisa Andosilla, Leila Gaudé, Marie-Noelle Lahouze-Davaud, Valérie Mirault, Georges Puissant et Jacques Koerber, ainsi que Serge Aeschlimann, Yves Reymond, Elvire Martinez et Micheline Vautravers.

Laure Fleury, Christiane Currat et Sandrine Giroud ont pris la peine de relire consciencieusement certaines parties de ce manuscrit et ont pu ainsi éviter que la plus grande partie de mes "petites" faiblesses en français n'apparaissent dans sa version finale.

Michel Blum a pris le relai de Nicolas dans le train, puisque nous avons effectué de nombreux Genève-Fribourg ensemble. Je le remercie particulièrement de m'avoir permis de partager son "spacieux appartement" fribourgeois pendant quelques mois.

Rosemarie et Max Matzinger m'ont chaleureusement accueilli chez eux, à Marly, pendant les derniers mois de la rédaction de ma thèse. Ils m'ont ainsi épargné de longues heures de voyage supplémentaires et permis de me concentrer pleinement sur mon travail. Je tiens également à remercier Chantal et Richard Pasquier, ainsi que Juliette et Laurent Excoffier, qui m'ont hébergé à l'occasion de mes nombreux séjours bernois. J'aimerais aussi remercier Françoise et Jean-Paul Giroud pour le vif intérêt qu'ils ont porté à mon travail et pour m'avoir permis d'en réaliser une partie sous le soleil de Sardaigne.

Le soutien constant de l'ensemble de ma famille, et plus particulièrement celui de Christiane, René, Déborah, Alexandre, Sandrine "Z", Didier, Maria, Irma, Gilbert et Lucienne a été indispensable à l'aboutissement de cette thèse.

Mes derniers remerciements vont bien évidemment à Sandrine, qui a été le complément nécessaire à l'achèvement de ce travail. Elle a toujours réussi à me motiver pendant les périodes de doute. Son œil avisé, à la fois interne au domaine scientifique et externe à la biologie, a été d'un apport inestimable. Je lui suis infiniment reconnaissant pour cela et pour tout le reste.

Table des matières

1 Introduction Générale

L'origine de notre espèce (*Homo sapiens sapiens*) et la reconstruction de son histoire sont des sujets fascinants qui ont toujours captivé les Hommes. Les moyens utilisés pour retracer le passé des humains n'ont cessé d'évoluer et diverses disciplines se sont attelées à cette tâche ardue. Parmi les plus récentes, la génétique des populations a été d'un apport inestimable à la compréhension de l'évolution d'*Homo sapiens sapiens*. Des approches permettant d'utiliser les données génétiques actuelles pour retracer l'évolution de l'Homme ont ainsi été développées et la complexité de ces techniques a rapidement augmenté au cours du temps. De même, le type de données génétiques étudiées a beaucoup évolué, passant des phénotypes (par exemple les configurations protéiques) aux génotypes (mutations portées par les séquences d'ADN)[1]. La génétique a permis d'aborder de façon complémentaire les problématiques proposées par les archéologues, les paléontologues et même par les linguistes. Le Laboratoire de Génétique et de Biométrie (LGB) de l'Université de Genève, dans lequel nous avons réalisé cette thèse, s'est notamment spécialisé dans une approche interdisciplinaire visant à étudier la variabilité génétique des populations humaines (Excoffier *et al.* 1987 ; Excoffier 1988 ; Sanchez-Mazas 1990 ; Dard *et al.* 1992 ; Currat 1999 ; Poloni 1999 ; Dard *et al.* 2001 ; Renquin *et al.* 2001 ; Buhler *et al.* 2002), ainsi que l'influence de la géographie et de la linguistique sur cette diversité (Excoffier *et al.* 1991 ; Poloni 1991 ; Dupanloup de Ceuninck 1999 ; Sanchez-Mazas 2000 ; Sagart *et al.* 2004).

La rapide avancée des techniques de laboratoire a permis, depuis une vingtaine d'années, la création de bases de données moléculaires utilisées pour retracer l'histoire de notre espèce. Ces données moléculaires ont notamment pu appuyer l'hypothèse d'une origine unique d'*Homo sapiens sapiens*, sans doute en Afrique – hypothèse connue sous le nom de "Out of Africa" (Stringer et Andrews 1988) – par oppostion à une origine multiple (ou multirégionale : Weidenreich 1946; Wolpoff 1989)[2]. Cette dernière théorie propose que l'évolution vers la forme finale de l'Homme moderne s'est faite de manière parallèle sur plusieurs continents. Il est cependant très difficile d'interpréter la structure génétique des populations actuelles en termes d'événements historiques ou préhistoriques et ces interprétations ne peuvent se faire qu'en étroite liaison avec les connaissances tirées d'autres sources comme l'archéologie ou la paléoanthropologie. En effet, si la variation des densités et des migrations influence fortement la structure génétique des populations, ce ne sont de loin pas les seuls facteurs impliqués. D'une part, le génome a son propre mode d'évolution, qui n'est encore que partiellement compris et qui peut être très variable en fonction des régions chromosomiques. D'autre part, des parties du génome peuvent être positivement ou négativement sélectionnées au cours du temps, cette sélection pouvant prendre des formes très variables et agir à des niveaux différents. Finalement, l'environnement joue un rôle prépondérant dans la mise en

---

[1] Voir par exemple Langaney (1988) pour une introduction sur de la diversité génétique des populations humaines et de leur étude.

[2] Voir Sanchez-Mazas (2001a, en français) ou Excoffier (2002, en anglais) pour une discussion à propos des différentes théories de l'origine de l'homme.

place de la structure génétique des populations, puisqu'il agit non seulement sur la sélection qui s'exerce sur leur génome, mais également très largement sur leur démographie et leur répartition. L'influence du milieu est d'autant plus importante que ses caractéristiques fluctuent en fonction des variations du climat. Ces dernières ont été extrêmement importantes pendant le dernier cycle glaciaire (120'000 dernières années) qui a vu l'apparition puis la diffusion d'*Homo sapiens sapiens*.

Il est donc difficile d'extraire la signature d'événements démographiques passés de la structure génétique des populations. De nombreuses recherches s'y consacrent pourtant, en inférant des hypothèses sur la démographie des populations à partir de données moléculaires (p.ex. : Mountain *et al.* 1995 ; Pritchard *et al.* 1999 ; Zhivotovsky *et al.* 2003). Il est cependant nécessaire d'avoir des modèles théoriques auxquels les données réelles peuvent être confrontées, afin de retenir les hypothèses les plus plausibles. De nombreux modèles analytiques ont donc été développés pour prédire les signatures génétiques attendues à la suite d'événements démographiques donnés, comme une croissance ou une contraction démographique ou comme le métissage de populations ou leur séparation. Ces modèles analytiques sont cependant limités, à la fois par la complexité des processus démographiques et par celle des données utilisées. En effet, si les génotypes comportent potentiellement plus d'informations que les phénotypes, leur complexité rend leur utilisation beaucoup plus difficile. La simulation de processus complexes offre donc une alternative prometteuse à leur compréhension lorsque leur résolution analytique est impossible. Elle permet, par exemple, de simuler les mouvements des individus constituant une population dans une aire virtuelle, en fonction de contraintes imposées par le modèle testé. Ces contraintes peuvent être des barrières géographiques, comme des montagnes ou des mers, qui empêchent la libre dispersion des individus. Ces derniers portent des gènes (eux-aussi virtuels) dont la constitution et la distribution sont comparées aux données réelles à la fin d'une simulation. La vraisemblance des données obtenues sous différentes hypothèses simulées peut ainsi être évaluée.

L'augmentation récente des capacités informatiques ouvre des perspectives immenses dans le domaine des simulations, puisqu'il est maintenant possible de prendre en compte la complexité des processus démographiques et génétiques dans un laps de temps raisonnable. C'est dans cette optique qu'a pris naissance le projet "Friction"[1] – dirigé par le Pr. Laurent Excoffier – dont le but était la reconstruction de l'histoire des populations humaines au moyen de données environnementales et génétiques. Notre thèse a été effectuée dans le cadre de ce projet. Ce type d'approche ayant encore été très peu exploré auparavant, la collecte d'informations et la réalisation d'un très grand nombre d'outils ont été nécessaires. Outre la compilation de données environnementales passées – réalisée par le Dr. Nicolas Ray – l'outil principal développé fut le logiciel de simulation SPLATCHE[2]. Ce programme, présenté au chapitre 2, permet de simuler à la fois la démographie et la génétique

---

[1] Le projet "Friction", attribué au Pr. Laurent Excoffier, a été financé par le Fond National Suisse pour la Recherche Scientifique, entre 1999 et 2003 (Fond n° 31-054059.98).

[2] "SPatiaL And Temporal Coalescences in Heterogeneous Environment", anciennement appelé "FRICTION", notamment dans Ray (2003).

d'une population évoluant dans une aire définie en utilisant au maximum la puissance informatique disponible. Comme l'a souligné Nicolas Ray (2003) – dont la thèse doit être considérée comme complémentaire à la nôtre – la grande difficulté du projet "Friction" a été de trouver le meilleur compromis entre un modèle suffisamment réaliste pour simuler de façon convaincante les processus désirés et un modèle suffisamment simple pour permettre leur compréhension. En effet, l'augmentation de la complexité d'un modèle n'est pas un gage de l'apport d'informations supplémentaires, puisque l'incertitude autour de nouveaux paramètres ne fait qu'augmenter celle qui existe autour des résultats obtenus. Même si notre programme a été développé afin d'étudier la dispersion des hommes modernes dans le monde, nous l'avons conçu de façon très générale, afin qu'il puisse être ultérieurement distribué à la communauté scientifique et qu'il permette d'aborder des questions diverses.

La préhistoire *d'Homo sapiens sapiens* en Europe nous a paru être un cadre particulièrement adapté à l'utilisation de SPLATCHE. D'une part, l'Europe est sans conteste le continent pour lequel l'histoire des populations humaines est la mieux connue, d'un point de vue archéologique mais également génétique. D'autre part, les variations climatiques qui ont affecté ce continent sont également abondamment documentées. Finalement, des hypothèses relativement bien définies ont été proposées pour expliquer l'histoire du peuplement humain de ce continent. Deux sujets d'études distincts ont rapidement émergé : premièrement, le remplacement des Néandertaliens par les premiers Hommes modernes lors de leur arrivée en Europe il y a environ 40'000 ans (Stringer et Andrews 1988) ; deuxièmement, le passage d'une économie de subsistance principalement basée sur la chasse, la pêche et la collecte de denrées sauvages, à une économie de production agricole. Cette transition, connue sous le nom de Néolithique (Lubbock 1865), a débuté au Proche-Orient il y a environ 10'000 ans (revue détaillée par Mazurié de Keroualin 2003). Nous proposons dans ce travail de fournir un cadre théorique à l'interprétation de la structure génétique européenne, en fonction des hypothèses émises pour ces deux événements démographiques majeurs. Les résultats de ces études sont présentés dans les chapitres 5 et 6, sous la forme de deux manuscrits soumis à publication.

Ces deux grandes périodes de transition démographique ont potentiellement pu laisser des traces dans la structure génétique actuelle des populations européennes. Elles coïncident avec la diffusion de nouvelles technologies dans l'ensemble du continent européen, à partir d'une petite région. Cette diffusion s'est vraisemblablement accompagnée d'importants mouvements de populations, qui peuvent être modélisés comme l'expansion spatiale d'une population à partir d'une source donnée. Cette expansion spatiale s'accompagne d'une croissance globale de la taille de la population. Si l'effet d'une simple croissance démographique sur la structure génétique des populations a déjà été passablement étudié, très peu d'informations sont disponibles sur l'influence de la diffusion spatiale d'une population en croissance. Dans ce travail, nous nous sommes donc intéressé à la composante spatiale de l'expansion d'une population et à ses effets sur la structure génétique. Dans le chapitre 3, nous avons tout d'abord étudié l'effet sur la diversité génétique d'une

expansion spatiale dans une aire inoccupée. L'effet d'une telle expansion est particulièrement intéressant, notamment dans les cas d'événements de spéciation ou de recolonisation post-glaciaire à partir d'une zone refuge.

Les deux événements démographiques qui nous préoccupent dans ce travail concernent deux populations distinctes: les Néandertaliens et les Hommes modernes dans un cas, les chasseurs-collecteurs[1] et les agriculteurs[2] dans l'autre. La version de base de SPLATCHE ne permettant pas de simuler simultanément deux populations différentes, nous avons donc dû procéder à des modifications qui sont décrites en détails dans l'ANNEXE 4. Parallèlement, il a fallu développer un modèle démographique qui permette de modéliser de manière réaliste les interactions entre deux populations, notamment la compétition et les échanges génétiques entre elles (chapitre 4). Lors des deux événements démographiques qui nous préoccupent ici, l'expansion spatiale d'une population invasive s'est faite dans une aire déjà occupée par une autre population. Nous avons donc testé les conséquences, dans les données moléculaires, de la diffusion d'une population dans un contexte occupé (chapitre 4). Avant cela, il a été nécessaire de cerner les valeurs les plus adéquates pour les différents paramètres du modèle, à partir des estimations faites pour les populations humaines contemporaines ou préhistoriques. Tous les développements nécessaires à la simulation de deux populations en interaction dans la même aire géographique, sont présentés dans le chapitre 4.

Il faut noter que les aspects techniques relatifs à l'implémentation du programme SPLATCHE ne sont mentionnés que succinctement dans ce manuscrit, bien que leur développement ait constitué une partie très importante de notre travail. De même, les nombreux outils complémentaires nécessaires à la manipulation et à l'extraction des innombrables données générées par nos simulations (parfois plusieurs centaines de milliers de fichiers différents) ne sont pas décrits dans ce manuscrit. Il s'agit en effet de "scripts Linux" et d'un programme en langage C++ ("WinReadSum") dont l'intérêt scientifique est très limité.

---

[1] Dans ce travail nous utiliserons préférentiellement le terme "**chasseurs-collecteurs**" au terme chasseurs-cueilleurs. Il est employé pour définir les individus appartenant aux populations dont le mode de subsistance est basé sur la chasse, la cueillette et la pêche. Nous ne ferons pas de distinction entre chasseurs-collecteurs mésolithiques et paléolithiques.

[2] Nous utiliserons le terme "**agriculteurs**" pour définir les individus néolithiques qui ont adopté la totalité des composantes définissant les sociétés agropastorales modernes, à savoir l'agriculture, l'élevage, la sédentarisation et la poterie (d'après Mazurié de Keroualin 2001).

2 Réalisation d'un logiciel permettant de simuler des données génétiques en fonction de paramètres démographiques et environnementaux.

2.1 Introduction

Comme nous l'avons déjà souligné dans l'introduction générale, de nombreuses études associent une structure génétique observée – notamment dans la population humaine – à un (ou des) événement(s) démographique(s) passé(s) (p. ex. : Menozzi *et al.* 1978; Sokal et Menozzi 1982 ; Piazza *et al.* 1995 ; Richards *et al.* 1996 ; Sajantila *et al.* 1996; Semino *et al.* 1996 ; Hammer *et al.* 1998 ; Torroni *et al.* 1998 ; Sykes 1999 ; Hill *et al.* 2000a ; Hammer *et al.* 2001 ; Helgason *et al.* 2001 ; Hurles *et al.* 2002; Capelli *et al.* 2003 ; Hurles *et al.* 2003 ; Quintana-Murci *et al.* 2003 ; Richards *et al.* 2003). En effet, la variation des densités des populations et les migrations influencent fortement la structure génétique des populations (Langaney *et al.* 1990; Barbujani *et al.* 1994 ; Lahr et Foley 1998; Stefan *et al.* 2001 ; Roebroeks 2003). En théorie, il est donc possible d'utiliser cette structure comme indice soutenant – ou infirmant – des hypothèses de peuplement proposées notamment par des disciplines comme l'archéologie ou la paléoanthropologie.

Malheureusement, l'interprétation de données génétiques pour inférer des informations démographiques est complexe, puisque de nombreux facteurs perturbateurs entrent en jeu (voir p. ex. : Langaney *et al.* 1992). Premièrement, des facteurs évolutifs intrinsèques au génome – comme la sélection naturelle et les effets de l'hétérogénéité des taux de mutation et de recombinaison – obscurcissent ou effacent la signature[1] génétique laissée par les événements démographiques passés (Lundstrom *et al.* 1992 ; Aris-Brosou et Excoffier 1996 ; Sanchez-Mazas 2001b ; Reich *et al.* 2002). Il faut donc être capable de reconnaître les effets de l'évolution du génome, puis de les séparer de ceux provoqués par l'histoire démographique d'une population. L'influence des facteurs évolutifs est encore, bien souvent, mal connue et il est difficile de s'en affranchir. Deuxièmement, l'histoire démographique des populations est rarement simple ; le passé des populations humaines est constitué d'une succession de processus complexes (migrations, contraction, expansion, etc…), qui se chevauchent souvent et dont l'importance est variable (Sokal 1991a ; Lahr et Foley 1998 ; Roebroeks 2003). Les interactions entre populations (affinité culturelle, compétition, assimilation) sont également déterminantes dans la mise en place de leur structure génétique (Sokal *et al.* 1993 ; Cappello *et al.* 1996 ; Sokal *et al.* 1996 ; Larruga *et al.* 2001). L'influence de l'environnement est

[1] Tout au long de ce travail, nous utiliserons le terme "**signature**" pour définir une structure génétique particulière qui résulte d'un scénario démographique donné. Il faut préciser que l'observation d'une telle structure dans les populations réelles peut être un indice en faveur de ce scénario mais ne constitute pas une preuve pour autant, puisque des structures génétiques similaires peuvent être générées par des processus différents (démographiques ou non, voir texte).

également prépondérante, puisque celui-ci joue un rôle non seulement sur les migrations (Brion *et al.* 2003) et les densités (Aborgast *et al.* 1996 ; Housley *et al.* 1997 ; de Menocal 2001), mais peut aussi agir comme facteur sélectif sur une partie du génome (Haldane 1949 ; Allison 1954; Sanchez-Mazas 2001b ; Currat *et al.* 2002). De plus, les caractéristiques environnementales évoluent au cours du temps, sous l'effet de la variation du climat (Adams et Faure 1997; Lahr et Foley 1998; Allen *et al.* 1999). Tous ces facteurs doivent donc être pris en compte lors de l'utilisation de données génétiques pour retracer l'histoire d'une espèce.

Malgré la complexité des processus évolutifs et démographiques, il est cependant possible d'effectuer des inférences sur l'histoire des populations à partir de données génétiques. En effet, des événements démographiques majeurs peuvent avoir laissé des traces (Menozzi *et al.* 1978; Sokal 1991b) observables avec un échantillonnage adéquat (Sokal et Jacquez 1991). Ces traces peuvent d'ailleurs être très différentes en fonction du type de données génétiques analysées (Kittles *et al.* 1999). Pour inférer un événement démographique par lequel est passée une population à partir de données génétiques, il faut connaître la signature génétique attendue après cet événement en dehors de toute influence perturbatrice. Il est en effet difficile d'utiliser une structure génétique observée pour soutenir une hypothèse de peuplement si l'on ne connaît pas la signature théorique attendue. Actuellement, des structures génétiques attendues pour des modèles simples sont connues, comme la signature laissée par une croissance démographique instantanée dans une population non-subdivisée (Tajima 1989b ; Slatkin et Hudson 1991; Rogers et Harpending 1992 ; King *et al.* 2000), la réduction de sa densité (Excoffier et Schneider 1999; Wahl *et al.* 2002) ou le métissage de plusieurs populations (Chikhi *et al.* 2001).

Ce chapitre est donc consacré au développement d'un outil informatique permettant d'étudier les conséquences d'un événement démographique sur la constitution génétique d'une ou de plusieurs populations. Cette approche vise à donner un cadre théorique à l'interprétation de données génétiques réelles en permettant la comparaison avec la structure génétique obtenue selon une hypothèse de peuplement donnée. Ce cadre n'est pas aussi précis que celui offert par des modèles analytiques, mais il a l'avantage de permettre la simulation de processus plus complexes, insolubles analytiquement. L'augmentation récente des capacités informatiques permet également de pousser beaucoup plus loin la complexité des modèles simulés, et de traiter une quantité d'informations plus importante. Cela permet également de tenir compte de la variabilité stochastique des processus génétiques au niveau du génome, mais également au niveau des populations.

La réalisation du logiciel de simulation "SPLATCHE" a été effectuée dans le cadre du projet "Friction". Ce projet vise à la reconstitution de l'histoire des migrations humaines, en fonction des données environnementales et génétiques. La complexité de ce projet a nécessité la participation de plusieurs personnes pendant près de 4 ans, chacune ayant un rôle bien défini. De nombreux aspects (logiciel SPLATCHE, compilation des données environnementales, modèles) ont, en effet,

dû être développés spécifiquement, puisqu'ils n'existaient pas avant sous la forme désirée. Le Pr. Laurent Excoffier est à la base même du projet et a principalement supervisé son développement. Le Dr. Nicolas Ray, s'est consacré à la recherche de données environnementales et à leur numérisation, ainsi qu'au développement et à l'implémentation des modèles démographiques. Notre propre rôle a principalement concerné l'incorporation de l'algorithme de simulation de données génétiques, ainsi que la visualisation de ses différentes composantes. D'autres personnes ont également contribué à l'avancée de ce projet. Le Dr. Stefan Schneider s'est consacré à la création de la structure initiale de SPLATCHE et le Pr. Jérôme Goudet a participé à l'élaboration de modèles démographiques.

Dès le début, SPLATCHE a été développé dans le but d'être utilisable de façon très générale, dans des cadres temporels et géographiques variables, afin d'étudier des processus démographiques complexes. Son intérêt réside dans sa capacité à traduire des données "écologiques" en données génétiques. SPLATCHE est donc un programme puissant qui permet de générer de nombreux types de données génétiques en incorporant de multiples paramètres démographiques et environnementaux. La réalisation de ce logiciel a été effectuée dans le cadre d'un vaste projet de recherche, dont les ramifications sont nombreuses. Il existe actuellement de nombreuses extensions de SPLATCHE – dont la version évolutive a gardé le nom de "FRICTION" – et leur nombre devrait encore augmenter dans le futur. Nous mentionnerons certains des développements en cours lorsque nous parlerons des perspectives de ce travail (chapitre 8). L'ANNEXE 3 présentera une version dérivée de SPLATCHE permettant la simulation de deux populations évoluant dans le même environnement.

Ce chapitre 2 présente, sous la forme d'un article publié dans *Molecular Ecology Notes* au début 2004 (section 2.2), les diverses possibilités offertes par SPLATCHE. La méthodologie sous-jacente à SPLATCHE ainsi que les données techniques sont présentées dans l'ANNEXE 2. Le manuel d'utilisation du logiciel constitue l'ANNEXE 1 et les possibilités de représentation graphique des généalogies de gènes l'ANNEXE 4.

2.2 Le Programme SPLATCHE

Notre article publié dans *Molecular Ecology Notes* présente le programme SPLATCHE et souligne ses applications potentielles. SPLATCHE est un logiciel qui permet de générer des données génétiques pour une population selon un scénario démographique donné. Ce scénario peut être conditionné par des informations environnementales, telles que la végétation, la topographie ou l'hydrographie. SPLATCHE se divise en deux parties: 1°) simulation de la démographie d'une population ; 2°) simulation de la structure génétique de cette population. Si la phase 1 peut être utilisée de manière indépendante, par exemple pour étudier les principales voies de migrations d'une espèce donnée, il n'en est pas de même pour la phase 2, qui ne peut se dérouler qu'à la suite de la première phase.

Les simulations se déroulent dans un monde virtuel défini par l'utilisateur à l'aide de cartes numériques représentant l'aire géographique d'intérêt. Cette aire géographique virtuelle est ensuite divisée en un certain nombre de cellules en fonction de la résolution désirée (Annexe 2.1). Il est ensuite possible de simuler la dispersion d'une population à partir d'une cellule source. Les densités et les migrations des individus appartenant à cette population sont conditionnées par les données environnementales propres à chaque cellule. Toutes les migrations d'individus, ainsi que l'évolution des densités de population à l'intérieur de chaque cellule au cours du temps, sont stockées dans une base de données. Il est ensuite possible d'extraire et de visualiser un grand nombre d'informations sur la démographie de la population à partir de cette base de données, notamment les directions de migrations préférentielles, ainsi que l'évolution des densités.

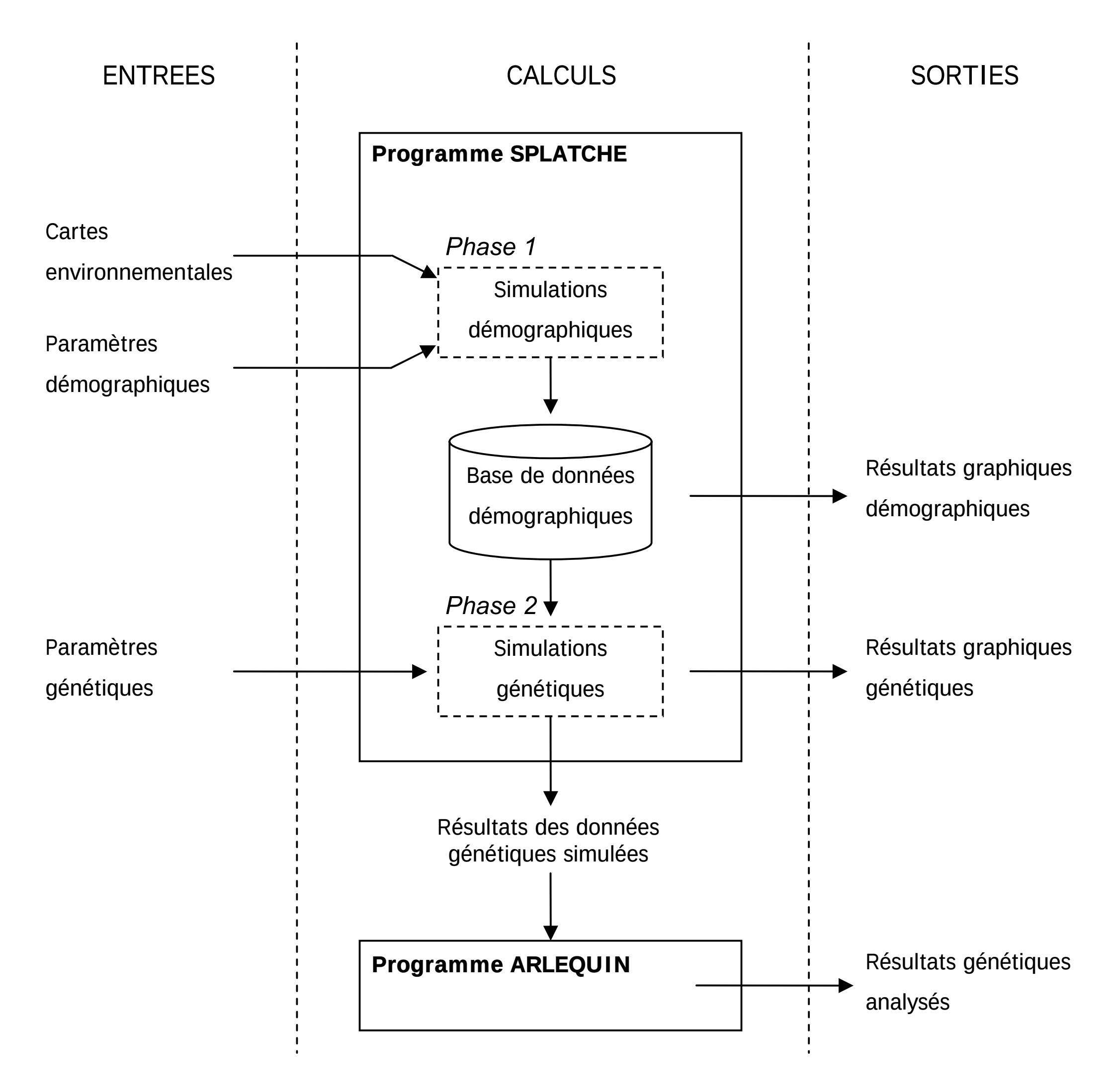


Figure 2.1. Schéma général des entrées, des calculs et des sorties liés au programme SPLATCHE.

A partir de la base de données démographiques, générée pendant la phase 1, il est possible de simuler des données génétiques pour un certain nombre d'échantillons tirés de la population

virtuelle. Le nombre et la localisation de ces échantillons sont définis par l'utilisateur, ainsi que le type de données générées (séquence d'ADN, microsatellite[1], fréquence allélique, RFLP[2]). La création de ces données génétiques se fait à l'aide de l'approche par coalescence (voir Annexe 2.2.1), qui permet la simulation de la diversité des gènes d'individus échantillonnés. Cette approche permet donc de réduire considérablement le temps de calcul, ainsi que l'espace mémoire nécessaires à la création de données génétiques car elle évite de simuler la diversité génétique de toute la population. L'économie ainsi faite permet de multiplier le nombre de simulations pour chaque scénario démographique et de tenir compte de la variablilité aléatoire du processus génétique. Il existe plusieurs formats de sortie pour les données génétiques. Premièrement, un format lisible par le logiciel ARLEQUIN (Schneider *et al.* 2000), qui permet d'analyser ces données. Deuxièmement, SPLATCHE permet la visualisation de certaines composantes spatiales de la structure génétique, notamment la distribution géographique des événements de coalescence et des "MRCA" (voir ANNEXE 3). Tous les détails concernant l'utilisation du programme SPLATCHE, ainsi que le type d'information générée sont présentés dans la section 2.2.1 et plus particulièrement dans l'ANNEXE 1[3].

Les utilisations potentielles de SPLATCHE sont nombreuses. Il est d'abord possible de l'utiliser pour étudier l'influence d'un événement démographique (expansion, contraction, migration, goulet d'étranglement, etc..) sur la diversité moléculaire (voir sections 3.2 et 3.3). Il est également possible d'utiliser SPLATCHE afin de comparer la structure génétique obtenue selon plusieurs hypothèses de peuplement aux données réelles, et ainsi de déterminer quelle est l'hypothèse la plus probable (Ray *et al.* 2004). Par ailleurs, SPLATCHE peut être utilisé pour prédire la dispersion spatiale d'une population en fonction de différents événements. SPLATCHE présente également un intérêt didactique important, puisqu'il permet de visualiser les composantes spatiales des généalogies de gènes (ANNEXE 3), et de faire directement la liaison entre leur topologie et l'information apportée par les données moléculaires. L'utilisation de SPLATCHE n'est bien évidemment pas restreinte à l'espèce humaine, puisque de nombreux types d'organismes différents peuvent être simulés, pour autant que leur écologie corresponde aux modèles démographiques proposés. En dernier lieu, SLATCHE est un logiciel évolutif, appelé à subir de nombreuses modifications dans le futur, en fonction des différentes applications pour lesquelles il pourra être utilisé.

2.2.1 Article

{ Page suivante }

[1] Les microsatellites (et STR) sont des séquences d'ADN de quelques paires de bases (1-6), qui sont répétées plusieurs fois à la suite. Ils présentent l'avantage d'être facilement amplifiés à l'aide d'une PCR. Leur mode d'évolution est encore relativement mal compris. Voir Zane *et al.* 2002 pour plus de détails.

[2] Les RFLPs (Restriction Fragment Length Polymorphism) sont de courtes séquences d'ADN (de 3 à 6) qui sont reconnues et coupées par des enzymes de restriction.

[3] Ces informations, ainsi que le programme lui-même, sont disponibles "on-line" à l'adresse www.cmpg.unibe.ch/software/splatche.

Molecular Ecology Notes (2004) **4**, 139–142 doi: 10.1046/j.1471-8286.2003.00582.x

PROGRAM NOTE

SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity

M. CURRAT,*†‡ N. RAY*†‡ and L. EXCOFFIER*
*Computational and Molecular Population Genetics Laboratory, Zoological Institute, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland, †Genetics and Biometry Laboratory, Anthropology and Ecology Department, University of Geneva, Rue Gustave Revilliod 12, 1227 Carouge, Switzerland

Abstract

We present a program called SPLATCHE (SPatiaL And Temporal Coalescences in Heterogeneous Environments) to simulate the molecular diversity of samples of genes in an environmentally heterogeneous world. Simulations are performed by, first, simulating the colonization of the world using environmental information to constrain migrations and local densities. These simulated densities and migration rates recorded over time and space are then used to simulate genetic diversity under a coalescent framework. The program thus virtually allows the translation of ecological information into molecular diversity, a novel approach that can be used to study the effect of climatic change on genetic diversity.

Keywords: coalescent simulation, demography, heterogeneous environment, molecular diversity, spatial expansion

Received 24 September 2003; revision accepted 20 November 2003

It is now widely admitted that climatic changes had (and still have) a profound impact on the distribution and the genetic diversity of many animal and plant species (e.g. Taberlet *et al*. 1998; Hewitt 2000; Barnes *et al*. 2002; Walther *et al*. 2002). At any given time, the distribution of a species and its diversity are heavily influenced by the environmental heterogeneity (e.g. Brachet *et al*. 1999; Wakeley & Aliacar 2001; Hanski & Ovaskainen 2003). For example, spatial fluctuations in food resources can lead to major differences in the sizes of subpopulations, and some landscape elements can act as barriers to migrations or in the contrary facilitate migration and act as corridors along which a larger amount of individuals can migrate. Therefore, due to the importance of the environment on real genetic processes, we have developed a program called SPLATCHE for the simulation of sampled molecular diversity, explicitly taking into account the spatial heterogeneity of the environment. SPLATCHE is actually based on a two-step simulation process: environmental information is first used to simulate (forward in time) the demographic and migration history of a set of subpopulations, and the resulting information is then used in a coalescent framework to generate (backward in time) the genealogy and the diversity of genes sampled at one or several locations (see Fig. 1). These two phases are enclosed in two separate modules of the program.

The simulation of the first demographic phase occurs in a two-dimensional stepping-stone (2DSS, Kimura & Weiss 1964), which is defined as an array of regularly spaced subpopulations or demes. Each deme can exchange migrants with its four neighbours. Information on local environments [such as vegetation or topographic maps, which can be imported from a Geographical Information System (GIS) package], is translated into two characteristics attached to each deme: a carrying capacity K and a friction value F. K represents here the maximum number of genes (or haploid individuals) that can be sustained by local resources, and F expresses the (relative) difficulty in moving through a deme, using relative values ranging from 0 (lowest friction, no barrier to migration) to 1 (highest friction, complete barrier to migration). Arbitrary levels of environmental heterogeneity can be considered. For instance, user-defined maps can be completely homogeneous, in term of K and F, but they can also reflect a more realistic world taking into account particular landscape features such as rivers, deserts, or mountains. These features can then be assigned specific F-values to reflect their potential roles as corridors

Correspondence: Professor Laurent Excoffier. Fax: + 41 31 631 48 88; E-mail: laurent.excoffier@zoo.unibe.ch
‡These two authors have contributed equally to this work.

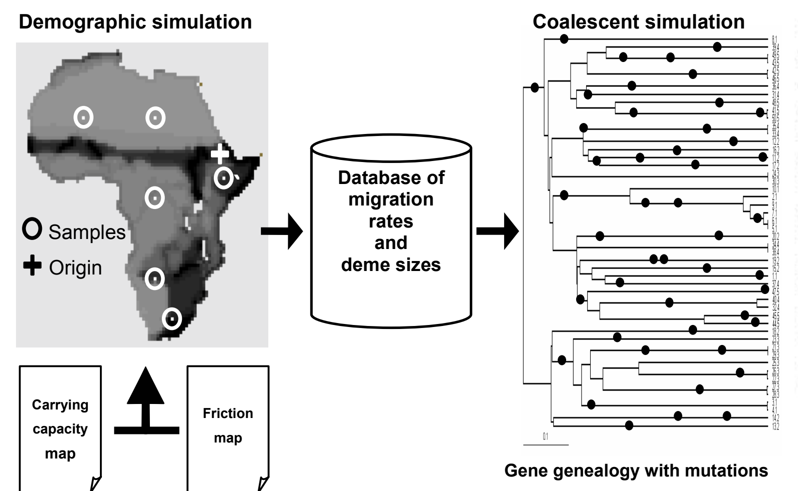


Fig. 1 Schematic view of the two-step simulation processes (demography and coalescent) necessary to generate genetic diversity at a given locus. The left pane illustrate the fact that forward demographic simulations are carried out from spatial information on potential carrying capacities and friction in different environment. The resulting simulated deme densities and immigration rates are used to perform coalescent simulations (right pane). Mutations following a given model are then sprayed on the coalescent tree to generate molecular diversity, and are shown as black dots on the gene genealogy.

or barriers to migration. Different K-values can also be assigned to reveal spatial differences in resource allocation. Different maps of K and F can be used at different times to simulate environmental changes. The maps to use and the time at which these changes occur are simply listed in input files. The format for the input maps is *ascii raster*, which can be generated by most GIS packages. A typical demographic simulation starts at one arbitrarily chosen deme, whose size and geographical location can be defined by the user in an input file. In each occupied deme, a growth phase is followed by a migration phase. The assumed growth model within deme is a standard logistic growth characterized by a user-defined intrinsic rate of growth per generation r, assumed constant over deme and over time, and a carrying capacity K depending on the environment. During the migration phase, the number of emigrants for every deme is computed as $N_t m$, where N_t is the local deme size at time t, and m is the migration rate. The number of emigrants can be allowed to vary stochastically as a Poisson variable with mean $N_t m$. The number of emigrants sent to each neighbouring deme is chosen from a multinomial distribution, with directional probabilities inversely proportional to the relative frictions of the neighbouring demes (see Ray 2003 for details). The sending of migrants to unoccupied demes results typically in a wave of advance of the whole population as shown on Fig. 2. The velocity and the shape of the edge of this wave depend on various parameters of the model, such as F, K, m, and r (see Ray 2003).

A modified version of the SIMCOAL program (Excoffier *et al*. 2000) has been integrated into SPLATCHE, and it is used to generate the molecular diversity of one or several samples. All former specifications of SIMCOAL are available except the 'historical events' formerly used to resize deme sizes and migration patterns between demes. Now, instead of manually setting the sizes of the demes and the pattern of migrations prevailing between demes at different time, this information is obtained from the previous demographic simulation round (see Fig. 1). The population size N_t of a given deme at time t is used to compute the probability of coalescence for every pair of genes found in that deme at time t. The number of emigrants (forward in time) from deme i to deme j at time t is used to compute the backward probability (backward in time) for a gene in deme j to migrate into deme i. SPLATCHE can generate restriction fragment length polymorphism, short tandem repeat (microsatellite data), DNA sequence data, and mere allele frequencies. The number and the spatial location of genes that must be simulated are specified in an input file.

In addition to text output in the ARLEQUIN format (Schneider *et al*. 2000), several graphical outputs are also available after a demographic simulation: deme size and number of migrants received from neighbouring demes through time, times of first colonization. Carrying capacities and frictions assigned to each deme can also be visualized as grayscale or colour bitmaps. They can alternatively be saved as *ascii raster* files to be imported into any GIS program. An innovative feature of SPLATCHE is the visualization of the spatial components of the coalescence trees, which can be translated into three different kinds of bitmaps: (i) the spatial distribution of the density of coalescence events, as shown for example in Ray *et al*. (2003); (ii) the genealogical connection between the nodes of the tree, in order to visualize the coalescent tree spatially onto the simulated world; (iii) the spatial distribution of the most recent common ancestor of all genes (MRCA), either for the whole tree or independently for each population sample. Two additional output files also allow one to obtain, for each sample, the spatial distribution of all coalescent events, and the time to the MRCA (T_{MRCA}).

The ability to translate environmental information into molecular diversity, via a demographic model, allows for a wide array of applications. Using a very simple homogeneous world, it has been possible to study the effect of a

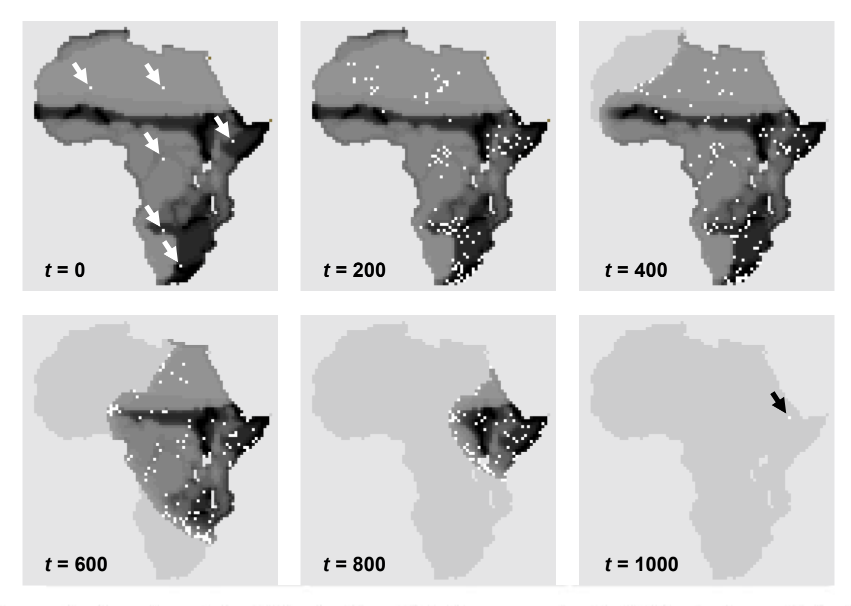


Fig. 2 A graphical example of a coalescent simulation in a hypothetical case assuming that Africa has been colonized from a single deme located in Eastern Africa (as shown by the black arrow on the lower right pane) 1000 generations ago. White arrows in the upper left pane indicate the locations of the sample demes at time zero, each containing 30 genes. The time *t* is expressed in number of generations from present, and goes backward in time. Darker demes have higher densities. At all times, white demes are those in which there is at least one gene lineage whose descendants can be traced forward to sampled genes. One sees that (going backward in time) gene lineages initially located in the sampled demes migrate to neighbouring demes and progressively diffuse in the environment. They are brought back to the origin of the settlement process by the shrinking of the occupied portion of the simulated world, which corresponds to the inverse of the assumed range expansion having started from East Africa.

range expansion on molecular diversity while varying parameters such as the carrying capacity and migration rates (Ray *et al*. 2003). Theoretical expectation of many statistics can be obtained with this simulation framework, because thousands of coalescent simulations can be generated in a reasonable amount of time, and later analysed with ARLEQUIN. The role of particular environmental or demographic variables can then be thoroughly assessed by analysing the molecular diversity, and obtaining the variance of the genetic processes. By using heterogeneous carrying capacity and friction maps, it is possible to consider very realistic situations in SPLATCHE. Specific metapopulation models could be implemented and SPLATCHE could also be used to make prediction about the spatial range and molecular diversity of a set of population after a change in the environment (man-mediated or following a climatic change). The graphical possibilities are also a major advantage of SPLATCHE. The visualized superposition of the demographic and coalescent processes can lead to a better understanding of some phenomenon, for instance in the study of the effect of migration corridor or spatial bottlenecks on genetic diversity, or the discovery of spatial regions that show an excess of migrants or of coalescent events. Finally, the educational value of SPLATCHE should be underlined, as many insights can be gained by visualizing the movements of genes through space and time in a realistic environment.

Executable Windows version of SPLATCHE, user guide and example files can be downloaded from http://cmpg.unibe.ch/software/splatche. Depending on the size of the simulated world and the number of generations, large amounts of RAM are required to guarantee reasonable execution time. For instance, a simulation of 10 000 demes over 4000 generations requires about 400 Mb of free RAM.

Acknowledgements

We are grateful to Stefan Schneider and Pierre Berthier for their computing assistance. This work was supported by a Swiss NSF grant no. 31-054059-98 to LE.

References

Barnes I, Matheus P, Shapiro B, Jensen D, Cooper A (2002) Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science*, **295**, 2267–2270.

Brachet S, Olivieri I, Godelle B *et al*. (1999) Dispersal and metapopulation viability in a heterogeneous landscape. *Journal of Theoretical Biology*, **198**, 479–495.

Excoffier L, Novembre J, Schneider S (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity*, **91**, 506–510.

Hanski I, Ovaskainen O (2003) Metapopulation theory for fragmented landscapes. *Theoretical Population Biology*, **64**, 119–127.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Kimura M, Weiss WH (1964) The stepping stone model of genetic structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561–576.

Ray N (2003) *Modélisation de la démographie des populations humaines préhistoriques à l'aide de données environnementales et génétiques*. PhD Thesis, University of Geneva, Switzerland.

Ray N, Currat M, Excoffier L (2003) Intra-deme molecular diversity in spatially expanding populations. *Molecular Biology and Evolution*, **20**, 76–86.

Schneider S, Roessli D, Excoffier L (2000) *ARLEQUIN: a software for population genetics data analysis*. User manual. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Geneva.

Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.

Wakeley J, Aliacar N (2001) Gene genealogies in a metapopulation. *Genetics*, **159**, 893–905.

Walther GR, Post E, Convey P *et al.* (2002) Ecological responses to recent climate change. *Nature*, **416**, 389–395.

2.3 Conclusion

Ce chapitre nous a permis de présenter le logiciel SPLATCHE, ainsi que certaines de ses applications potentielles. Il s'agit d'un logiciel généraliste, permettant de simuler une grande variété de scénarios démographiques et de données génétiques associées. Il permet de visualiser de nombreux aspects de la généalogie d'un échantillon de gènes tirés d'une population (voir ANNEXE 3). Ce logiciel offre un grand intérêt didactique puisqu'il permet la compréhension des relations entre la démographie d'une population et sa diversité moléculaire. De plus, il offre de nombreuses perspectives que nous discuterons de manière plus détaillée dans la conclusion finale de ce travail (chapitre 8). Ce logiciel a été mis à la disposition de la communauté scientifique, par l'intermédiaire d'un site web (http://cmpg.unibe.ch/software/splatche).

Bien que des extensions de la théorie de la coalescence aient été développées dans le cas de populations subdivisées (Notohara 1990 ; Marjoram et Donnelly 1994 ; Slatkin 1995 ; Rousset 1996 ; Wakeley 1999 ; 2000, 2001; Wakeley et Aliacar 2001), la résolution analytique de ces modèles devient problématique lorsque la complexité des situations considérées augmente. SPLATCHE propose donc une alternative en offrant la possibilité de simuler des données génétiques pour des situations démographiques relativement complexes. Dans le chapitre 3, nous présentons deux applications du logiciel SPLATCHE dans un contexte relativement complexe. Puis, dans les chapitres 5 et 6, nous présentons également deux recherches effectuées à l'aide d'une version dérivée du programme SPLATCHE.

3 Expansion spatiale dans un contexte inoccupé

3.1 Introduction

Les expansions spatiales de populations ont été fréquentes pendant le Quaternaire, non seulement pour l'espèce humaine, mais également pour de nombreux autres organismes. En effet, lors des périodes glaciaires, la répartition géographique de nombreuses espèces se réduit à des zones refuges de petite taille (voir p. ex. : Taberlet *et al.* 1998 ; Hewitt 2000). Lorsque le climat devient plus clément, ces espèces – qui ont passé plusieurs générations avec des effectifs faibles – colonisent de nouveaux territoires à partir des zones refuges. Il s'ensuit donc une expansion spatiale combinée à une croissance démographique, souvent dans des zones inoccupées par le même type d'organisme. Les conséquences d'une expansion spatiale sur la structure génétique d'une population ont été encore très peu étudiées et nous nous y intéressons ici.

Dans ce chapitre nous exposons deux applications du logiciel SPLATCHE, lui-même présenté dans le chapitre 2. La première, qui est présentée sous la forme d'un article publiée dans *Molecular Biology and Evolution* en 2003, concerne l'étude de la diversité génétique intrapopulationnelle observée après une expansion spatiale et démographique (section 3.2). Cette diversité est simulée sous la forme de séquences d'ADN semblables à celles étudiées pour le génome mitochondrial[1] chez l'Homme. La seconde étude est complémentaire à la première puisqu'elle étudie les effets du même genre d'expansion, mais cette fois sur un autre type de marqueur moléculaire : les SNPs[2] (section 3.3). Les SNPs ont en effet été passablemet typés sur le pendant masculin du génome mitochondrial : la portion non-recombinante du chromosome Y, ou MSY[3]. Nous discutons ensuite les observations faites à l'aide des deux types de marqueurs dans le contexte des populations européennes (3.4).

Avant d'aller plus loin, il est important de définir deux termes que nous utiliserons par la suite. Dans ce travail, le terme **population** signifie : "*un groupe d'individus, qui, à un moment donné dans*

---

[1] Le génome mitochondrial est transmis uniquement pas voie maternelle. Il est long d'environ 16 kb et contient notamment deux régions hypervariables (HV1 et HV2) - localisées dans la région non codante (D-loop) - qui ont été complètement séquencées dans de nombreuses populations humaines depuis 1981 (Anderson *et al.* 1981). Sa présence en grande quantité dans une cellule (1'000 à 10'000 fois l'ADN nucléaire) et son haploïdie font de l'ADN mitochondrial un sujet d'étude très important puisqu'il peut être séquencé directement et qu'il évolue beaucoup plus rapidement que l'ADN nucléaire.

[2] Les SNPs (Single Nucleotide Polymorphism) - aussi appelés UEPs, pour Unique Event Polymorphism(s) - sont des locus polymorphes pour lesquel on connaît deux états : présent ou absent. Ils sont souvent utilisés pour étudier de larges portions du génome. Voir par exemple Vignal *et al.* 2002 et Shastry 2002 pour plus de détails.

[3] Le chromosome Y est transmis uniquement de père en fils et est le plus petit chromosome du génome humain. Il est néanmoins environ 4'000 fois plus grand que le génome mitochondrial puisqu'il est constitué de près de 60 millions de paires de bases. Sa majeure partie (95%) ne recombine pas et est appelée MSY (anciennement NRY, voir Skaletsky *et al.* 2003). Cette portion du chromosome est spécifique aux mâles (chez les mammifères) et permet la détermination génétique du sexe. Elle est entourée de deux régions recombinantes, appelées "régions pseudo-autosomales".

le temps, partagent au moins une caractéristique définie par le chercheur" (Forster *et al.* 2002). Cette caractéristique peut être biologique, géographique ou culturelle. Dans ce travail, une population peut notamment être synonyme de communauté, par exemple dans le cas de la simulation des interactions entre chasseurs-collecteurs et agriculteurs. Dans ce manuscrit, le terme **dème** (Gilmour et Gregor 1939) fait référence à un groupe d'individus appartenant à la même population (par exemple à la même communauté selon la définition ci-dessus) et à la même aire géographique (représentée dans SPLATCHE par une cellule virtuelle, voir page 173).

3.2 Diversité moléculaire intrapopulationnelle à la suite d'une expansion spatiale

La signature génétique laissée par une population dont la taille est passée d'un très petit à un très grand nombre d'individus a été passablement étudiée. Slatkin et Hudson (1991, Figure 9.7A, dans l'ANNEXE 2) ont notamment montré qu'une expansion démographique provoque une généalogie de gènes en forme de peigne ("star-like"). Il en résulte une distribution "mismatch"[1] unimodale avec une forme en cloche, alors que, dans une population stationnaire, la distribution "mismatch" est multimodale (Rogers et Harpending 1992, Figure 9.7C). Bien que l'expansion spatiale d'une population conduise également à un accroissement démographique global, la ressemblance entre la signature génétique obtenue dans cette situation et celle obtenue dans le cas d'une simple croissance démographique dans une population non-subdivisée était inconnue. Très peu d'attention avait en effet été accordée à l'influence de la dispersion spatiale d'une population en expansion démographique. Nous nous sommes donc intéressé à cette question. Dans un article publié en 2003, dans la revue *Molecular Biology and Evolution*, nous décrivons la diversité moléculaire attendue dans un échantillon issu d'une population ayant passé par une expansion démographique et spatiale. A l'aide du programme SPLATCHE, nous simulons la diffusion spatiale d'une population dans une matrice de dèmes homogènes, à partir d'un seul dème source, selon différents paramètres démographiques. Le but de ces simulations est d'étudier la signature génétique observée dans une population subdivisée en expansion, et de la comparer avec celle attendue après une croissance démographique dans une population non-subdivisée.

Dans cette étude, nous montrons que la généalogie d'une population ne dépend pas seulement de l'âge de son expansion, mais également du flux génique qui existe entre les dèmes qui la constituent. Ce flux migratoire est mesuré par le produit *Nm*, qui est le nombre de migrants échangés entre dèmes – lorsque l'équilibre démographique est atteint – où *N* correspond à la densité d'un dème et *m* à la proportion de migrants échangés entre dèmes voisins. Deux types de signatures génétiques distincts sont observés en fonction de l'importance du flux migratoire. Lorsque *Nm* est faible (moins de 20 individus échangés entre dèmes voisins à l'équilibre), de

[1] La distribution "mismatch" correspond à la distribution du nombre de différences entre paire de séquences d'ADN provenant d'une population donnée.

nombreuses coalescences[1] sont très récentes et se déroulent dans la région dans laquelle a eu lieu l'échantillonnage (article 3.2.1 : Figure 1). Cela reflète des liens de parenté récents entre les individus échantillonnés. A l'inverse, lorsque *Nm* est grand, la majorité des coalescences se déroulent dans la région de la source de l'expansion, au moment du début de celle-ci (article 3.2.1 : Figure 1). Les liens de parenté entre la plupart des individus tirés d'un échantillon remontent alors à l'origine de la population.

Au niveau des généalogies de gènes et des données moléculaires, un grand *Nm* se traduit par des arbres en peigne avec de longues branches terminales (article 3.2.1 : Figure 2), et, par conséquent, par des distributions "mismatch" unimodales et une faible variance (article 3.2.1 : Figure 3). Un petit *Nm* donne lieu à une combinaison de courtes et de longues branches terminales et, par conséquent, à des distributions "mismatch" multimodales et une grande variance. Les statistiques utilisées habituellement pour détecter une expansion démographique, telles que *D* (Tajima 1989a, 1989b) et *Fs* (Fu 1997) ne sont efficaces que lorsque le *Nm* associé aux dèmes est grand (article 3.2.1 : Table 1). Même si une population est passée par une expansion spatiale, *D* et *Fs* la détectent très difficilement lorsque le *Nm* de cette population est faible.

Nous montrons donc que la croissance démographique d'une population combinée à une expansion spatiale n'implique pas toujours la même signature moléculaire qu'une simple croissance démographique dans une population non-subdvisée. Les deux signatures sont identiques uniquement si le flux génique entre les dèmes qui constituent la population subdivisée est grand.

Ces observations permettent d'expliquer, par une simple différence de densité, pourquoi les distributions "mismatch" obtenues pour le génome mitochondrial humain sont unimodales dans les populations post-néolithiques, et multimodales dans les populations de chasseurs-collecteurs (Watson *et al.* 1996 ; Excoffier et Schneider 1999, voir aussi la Figure 6 de l'article 3.2.1). Il n'est donc pas nécessaire d'invoquer une réduction de la taille des populations de chasseurs-collecteurs lors du Néolithique (Excoffier et Schneider 1999) pour expliquer leurs distributions "mismatch" multimodales.

Finalement, le fait que la diversité moléculaire observée dans les populations actuelles dépende du flux migratoire récent entre sous-populations suggère qu'il doit être possible d'estimer le produit *Nm* d'une population à partir d'un seul des dèmes qui la composent. Cette perspective a d'ailleurs donné lieu à une publication (Excoffier 2004).

3.2.1 Article

{ Page suivante }

---

[1] Se référer à l'Annexe 2.2.1 pour une description du processus de coalescence et des définitions qui s'y rattachent.

Intra-Deme Molecular Diversity in Spatially Expanding Populations

Nicolas Ray,† Mathias Currat,*† and Laurent Excoffier†*

*Genetics and Biometry Lab, Department of Anthropology and Ecology, University of Geneva, Geneva, Switzerland; and †Computational and Molecular Population Genetics Lab, Zoological Institute, University of Bern, Bern, Switzerland

We report here a simulation study examining the effect of a recent spatial expansion on the pattern of molecular diversity within a deme. We first simulate a range expansion in a virtual world consisting in a two-dimensional array of demes exchanging a given proportion of migrants (*m*) with their neighbors. The recorded demographic and migration histories are then used under a coalescent approach to generate the genetic diversity in a sample of genes. We find that the shape of the gene genealogies and the overall pattern of diversity within demes depend not only on the age of the expansion but also on the level of gene flow between neighboring demes, as measured by the product *Nm*, where *N* is the size of a deme. For small *Nm* values (< approximately 20 migrants sent outwards per generation), a substantial proportion of coalescent events occur early in the genealogy, whereas with larger levels of gene flow, most coalescent events occur around the time of the onset of the spatial expansion. Gene genealogies are star shaped, and mismatch distributions are unimodal after a range expansion for large *Nm* values. In contrast, gene genealogies present a mixture of both very short and very long branch lengths, and mismatch distributions are multimodal for small *Nm* values. It follows that statistics used in tests of selective neutrality like Tajima's *D* statistic or Fu's F_S statistic will show very significant negative values after a spatial expansion only in demes with high *Nm* values. In the context of human evolution, this difference could explain very simply the fact that analyses of samples of mitochondrial DNA sequences reveal multimodal mismatch distributions in hunter-gatherers and unimodal distributions in post-Neolithic populations. Indeed, the current simulations show that a recent increase in deme size (resulting in a larger *Nm* value) is sufficient to prevent recent coalescent events and thus lead to unimodal mismatch distributions, even if deme sizes (and therefore *Nm* values) were previously much smaller. The fact that molecular diversity within deme is so dependent on recent levels of gene flow suggests that it should be possible to estimate *Nm* values from samples drawn from a single deme.

Introduction

The connection between the past history of a population and its neutral genetic diversity has become obvious with the advent of coalescent theory (Kingman 1982*a*, 1982*b*; Hudson 1990; Nordborg 2001). Although coalescent theory was initially developed in the context of a single population, it has been rapidly extended to include subdivided populations or populations connected by migration (the structured coalescent) (Notohara 1990; Marjoram and Donnelly 1994; Slatkin 1995; Rousset 1996, 1997; Nordborg 1997; Wilkinson-Herbots 1998; Wakeley 1999, 2000, 2001; Wakeley and Aliacar 2001). Past theoretical studies have focused on island or stepping-stone models within homogeneous environments. Focusing on a finite island model, Wakeley (1999, 2001) has shown that the coalescent process in a subdivided population could be divided into two distinct phases when the number of demes is large (much larger than the number of sampled genes). Going backward in time, the first phase, called the "scattering phase," is usually rapid and ends when all sampled genes are located in different demes. It is characterized by a series of initial coalescent events, with migration events scattering the gene lineages into different demes. The second phase, called the "collecting phase," is usually much longer and describes the coalescent process between the end of the scattering phase and the ultimate coalescent event. This phase is characterized by a large number of migration events and a few coalescent events that are only possible when a gene lineage has migrated into a deme already occupied by another gene lineage. Interestingly, the coalescent during the collecting phase is similar to that of an unsubdivided population but on a timescale proportional to the effective size of the whole population, itself depending on the number of demes, the migration rate between demes, and the deme size (Wakeley 1999). Additional realism has been recently incorporated by allowing for the occurrence of a potentially changing number of demes of unequal size connected by potentially changing rates of migration (Wakeley 2001; Wakeley and Aliacar 2001), showing that coalescent events would accumulate over time in small demes with low migrations rates (Wakeley 2001). Coalescent-based approaches have also been developed to estimate nonhomogeneous and asymmetric migration rates among demes of unequal sizes (Beerli and Felsenstein 1999, 2001), albeit under the assumption that the sampled demes actually exchange migrants.

The development of more realistic models that incorporate demographic history may allow for the explanation of complex patterns that may be apparent in population genetic data. A classical example of the influence of the demographic history of a population on its molecular diversity is a recent demographic expansion that leads to starlike phylogenies (Slatkin and Hudson 1991) and to unimodal distributions of the number of pairwise difference or mismatch distributions (Rogers and Harpending 1992). While this pattern could also be obtained by a complex mutation mechanism in the absence of large expansions, for instance, heterogeneity of mutation rates (Lundstrom, Taravé, and Ward 1992; Aris-Brosou and Excoffier 1996), the study of mitochondrial DNA in many human populations suggests that most human populations have experienced Pleistocene demographic expansions (Sherry et al. 1994; Rogers

Key words: mismatch distribution, spatial expansion, demographic expansion, human evolution, mitochondrial DNA, subdivided population.

E-mail: laurent.excoffier@zoo.unibe.ch.

Mol. Biol. Evol. 20(1):76–86. 2003
DOI: 10.1093/molbev/msg009

1995; Rogers and Jorde 1995; Harpending et al. 1998; Excoffier and Schneider 1999; Schneider and Excoffier 1999). Similarly, microsatellite data from the Y chromosome were better explained with models based on past expansion than on stationarity (Pritchard et al. 1999). In contrast, analyses of Y chromosome single nucleotide polymorphism (SNP) did not provide any clear evidence for demographic expansions (Pereira et al. 2001). Studies with nuclear markers have also provided ambiguous results. Signals of expansion were found in some but not in all populations analyzed for microsatellite data (Reich and Goldstein 1998; Beaumont 1999; Goldstein et al. 1999). SNP studies showed no signs of expansion when single populations were considered (Nielsen 2000; Wakeley et al. 2001), whereas signals of expansions were found in a subdivided population model (Wakeley et al. 2001).

It is apparent that under existing demographic models, it is difficult to establish a clear and consistent explanation for the observed patterns of human molecular diversity. Discrepancies regarding signs of demographic expansions may be due to differences in demographic histories among regions (Reich and Goldstein 1998; Goldstein et al. 1999) and among ethnic groups (food producers vs. food gatherers) (Watson et al. 1996; Excoffier and Schneider 1999), differences between loci (Beaumont 1999), ascertainment bias in the choice of markers (Wakeley et al. 2001), or a lack of resolution of some markers (Pereira et al. 2001). However, these discrepancies could also result from making inferences based on erroneous models of population history (e.g., if the population is indeed subdivided) (Marjoram and Donnelly 1994).

While extensive studies have focused on the effect of population subdivision on the shape of gene genealogies (e.g., Notohara 1990; Marjoram and Donnelly 1994; Donnelly and Tavaré 1995; Nordborg 1997; Wakeley 2001), the effect of range or spatial expansions have thus far been neglected. In the case of modern humans, estimations of the age of the demographic expansions obtained from mtDNA sequence analyses point to the Pleistocene, and so these expansions could indeed represent a global increase in effective population size due to the spread of humans after a bottleneck. Although previous work has suggested that observed patterns of molecular diversity may have resulted from a simple demographic increase, the possibility also exists that these patterns are a signal of a range expansion after a speciation event (Excoffier and Schneider 2000). Although a range expansion certainly leads to an increase in the global effective size of a species, it is not known whether it leads to exactly the same molecular signal as a demographic expansion in a single unsubdivided population.

Despite advances in analytical techniques that allow for estimates of population parameters in more realistic settings, they may become intractable under complex evolutionary scenarios. It appears, therefore, that coalescent simulations are still useful and necessary to investigate the effect of such complex scenarios (such as nonconstant environments) on various aspects of the molecular diversity of populations. In this study, we use a simulation framework to study the combined effect of spatial and demographic expansions on patterns of within-deme molecular diversity in a simple two-dimensional landscape. After simulating a wave of advance (using a simple migration model with logistic regulation of deme size), a coalescent approach is used to simulate the genetic diversity of a sample conditional on the demographic history of the population. Different aspects of the molecular diversity are recorded, and factors with the potential to affect molecular diversity (place of origin, local deme size, size of gene flow between neighboring demes, and sampling location) are investigated and discussed.

Material and Methods

To efficiently simulate the molecular diversity expected in a sample drawn from a deme belonging to a large subdivided population having gone through a recent spatial expansion, we proceed in two steps. We first use a forward simulation scheme to generate the demography (density and gene flow) of a two-dimensional array of demes initially empty except for a single deme assumed to be at carrying capacity. We then use this resulting demographic information to simulate the molecular diversity of a set of DNA sequences drawn from a single deme using a coalescent backward approach.

Demographic Simulations

Simulations were performed in a subdivided population consisting of 2,500 demes arranged in a two-dimensional stepping-stone lattice of 50×50 demes. At the beginning of a simulation, a single deme of this population is occupied with a density equal to 100 (unless specified otherwise). This ancestral deme is the source of an isotropic spatial expansion. In our simulations, we have considered just two potential locations for this ancestral deme: one was located at the center of the lattice (at position $<25; 25>$), and the other located near the periphery (at position $<5; 5>$). After the onset of the spatial expansion process, the range of the population increases due to ongoing exchange of migrants between occupied demes and their neighbors. Emigrants are sent from a given deme having density N_t at time t to neighboring demes at rate m, so that $N_t m$ emigrants are sent outwards at each generation. If a gene is sent to an occupied deme, the movement results in gene flow. If not, the movement results in the colonization of a new deme. The emigration rate does not depend on the current density of the target deme, so that the same proportion of migrants are sent to empty or occupied demes. The number of emigrants $N_t m$ is then distributed equally among the neighboring demes. The density of each deme is limited by its carrying capacity K, and is regulated logistically as

$$N_{t+1} = N_t[1 + r([K - N_t]/K)],$$

where N_t is its density at time t, and r is the intrinsic rate of increase per generation (in the current study, r was constant at 0.1). At carrying capacity, Km migrants are thus exchanged between a deme and its neighbors. In the following, this number of migrants exchanged at equilib-

rium will be denoted by Nm to be consistent with published literature. For each generation, we implement a logistic regulation step followed by a round of migration. The demographic simulations are performed for 4,000 generations, and we store for each generation t the density of the j-th deme (N_{jt}) and the number of immigrants received from the k-th deme (I_{jkt}) in a database. This demographic database is then used to perform the genetic simulations using a coalescent approach described below.

Coalescent Simulations

Under neutrality, the genetic diversity of samples in a subdivided population is easy to simulate, as it depends only on the demographic and migration histories of the demes (e.g., Hudson 1990; Nordborg 2001). For this purpose, we have modified the coalescent simulation program SIMCOAL (Excoffier, Novembre, and Schneider 2000), allowing it to take into account the dynamic nature of deme densities and migration rates between adjacent demes. Starting at the present generation, we simulate the genealogy of genes sampled in a deme located, for convenience, at one of the two previously specified positions in the grid. Because we are interested in describing intra-deme diversity, we stress the fact that samples of genes are always drawn from a single deme. At each generation and going backward in time, genes can either move to a different deme or coalesce if they are not the single gene lineage in their deme. At generation t, the probability of emigration of a gene from deme j to deme k is computed according to the information recorded in the database created during the demographic simulation step and is equal to I_{jkt}/N_{jt}. After migration, the probability of a coalescence event in deme j depends both on the number of genes (i) present in deme j and on its density at time t as $i(i-1)/(2N_{jt})$. For each generation, we first implement a coalescence phase followed by a migration phase. As usually assumed in analytical treatments, a single coalescent event is allowed per deme per generation. In the case where the deme size is not much larger than the number of gene lineages (i) present in that deme, this strategy leads to slightly longer coalescence times (up to i generations) than if several coalescent events were allowed per generation. Because i is smaller than 30 in our current simulations, it is unlikely to affect the pattern of molecular diversity that is generated over thousands of generations. The coalescent process stops when there is a single gene lineage left in the array of demes. In the case when multiple gene lineages trace back to the ancestral deme at a time corresponding to the beginning of the forward simulation, the backward coalescent process proceeds further in this single deme of density equal to its initial density (100, unless specified otherwise). During the simulations, we record the locations and times of all coalescent events. For each simulated gene genealogy, we simulate mutations on the branches of the genealogy according to a Poisson process with rate μt, where μ is the mutation rate and t is the length (in generations) of a given branch. In the present case, we simulated an unbiased substitution process on a sequence of DNA of 300 bp, with $\mu = 0.001$ for the whole sequence, assuming a finite-site mutation model without heterogeneity of mutation rates. One thousand coalescent simulations were performed for each set of demographic parameters tested.

The distribution of a number of statistics were gathered from the simulated samples, including the number of segregating sites (S), the average number of pairwise differences (π), Tajima's D statistic (Tajima 1989), Fu's F_S statistic (Fu 1997), and the mismatch distribution. All analyses were performed using the software ARLEQUIN (Schneider, Roessli, and Excoffier 2000). Unless specified otherwise, summary statistics and mismatch distributions were obtained from the simulation of samples of 30 DNA sequences.

Results

Spatial and Temporal Distribution of Coalescent Events

In figure 1, we show various aspects of the dynamics of the spatial expansion process for two different numbers of migrants (Nm) exchanged between neighboring demes. The first obvious result (fig. 1*A* and 1*B*) is that for low Nm values ($Nm = 10$), the speed of the colonization wave is slower (600 generations to colonize the 2,500 demes) than with high Nm values ($Nm = 500$) (400 generations to colonize the 2,500 demes). Note that this effect is not due to a difference in the proportion of migrants, since in both cases m was set to 0.1. This is rather due to the fact that for low Nm values, the deme takes longer to fill than for higher Nm values, and therefore migration commences later. The migration pattern also influences the timing and the location of coalescent events. A majority of coalescent events are recent, having occurred during the scattering phase (*sensu* Wakeley 1999) and been geographically located close to the sampling location for $Nm = 10$ (fig. 1*C* and *E*), while for $Nm = 500$, they are mainly older and located close to the origin of the expansion (fig. 1*D* and *F*). Note that the ultimate coalescent event is older than 4,000 generations in 96.1% of the cases (and thus occurs in the ancestral deme) when $Nm = 10$, compared with 100% of the simulated cases when $Nm = 500$.

Patterns of Molecular Diversity

We have studied different aspects of DNA sequence polymorphism within a single deme drawn from a population that has experienced a range expansion. The results of the analysis of simulated samples are reported in table 1 for different levels of migration among neighboring demes. A drastic difference is found between demes exchanging 20 migrants or less per generation and demes exchanging higher numbers of migrants. Whereas the average number of pairwise differences only slightly increases with larger Nm values (going from $\pi = 6.3$ for $Nm = 5$ to $\pi = 8$ for Nm values ≥ 200), the number of segregating sites increases much more drastically (going from $S = 30$ for $Nm = 5$ to $S = 96.9$ for $Nm = 1{,}000$). This difference can be attributed to the timing of the coalescent events, which is indeed different for small or large Nm values. Because a majority of coalescent events occur in the scattering phase for small Nm values and much later (around the onset of the expansion) for large

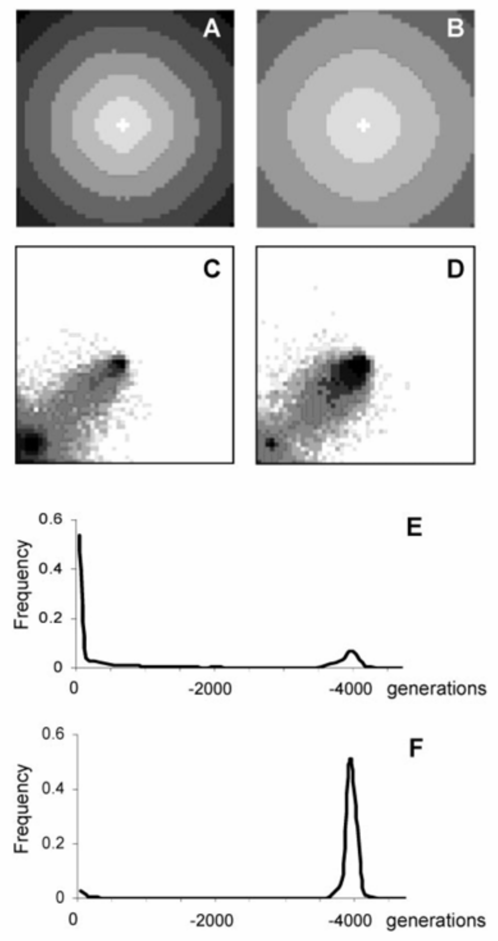


FIG. 1.—Summary of the dynamics of the spatial expansion process and its associated coalescent in a simulated subdivided population of 2,500 demes arranged as a two-dimensional stepping-stone (50 × 50 demes). *A and B*, Dynamics of the spatial expansion showing the progressive colonization of demes, with the spatial expansion starting from the central deme. Each shade of gray denotes the limit of the area of occupied demes after a further 100-generation step. The central white "cross" is the state of the expansion after just one generation. *C and D*, Empirical spatial distribution of the coalescence events obtained from 1,000 simulations of the genealogies of 30 genes sampled in a single deme in the lower left periphery (deme at position <5; 5>) of the population. Gray intensity is related to the total number of coalescent events having occurred in a given cell. Cells where no coalescent events occurred are shown in white. *E and F*, Empirical time distribution of the coalescence events obtained from 1,000 simulations. Time before present is represented on the X axis. In *A*, *C*, and E, the number of migrants exchanged between neighboring demes is lower ($Nm = 10$) than in *B*, *D*, and *F* ($Nm = 500$). In all cases, m is set to 0.1.

Nm values (see fig. 1*E* and *F*), the total length of the gene genealogy is much larger for samples of genes drawn from a deme with high *Nm* than those drawn from demes with small *Nm*. The difference in the timing of coalescence events and the overall shape and length of genealogies of genes drawn from demes with low or high *Nm* values can be seen in figure 2, where we show three random genealogies for three values of *Nm* (5, 25, and 200). As could be inferred from figure 1, there are many recent coalescent events in demes sending only a few migrants to neighboring demes, whereas recent coalescent events are rare in demes sending many migrants, resulting in very long terminal branches in the genealogies. Note that very similar gene genealogies with long terminal branches are observed in an unsubdivided population after a demographic expansion (Slatkin and Hudson 1991; Rogers and Harpending 1992).

The sampling location and the geographical location of the expansion have no effect on the pattern of molecular diversity in our homogeneous environment for large *Nm* values, whereas we observe a slight reduction in genetic diversity for demes that are sampled in the periphery of the simulated population for low *Nm* values (independent of the origin of the expansion) (table 2). Tajima's *D* statistic seems sensitive to the sampling location for low *Nm* values, as demes sampled in the center show a significant negative *D* value in 22% to 26 % of the cases at the 5% level, whereas demes sampled in the periphery only show significant *D* values in 7% to 8% of the simulations.

Mismatch Distributions

The empirical distributions of the number of differences between pairs of genes (identified here as mismatch distributions for sake of brevity) are shown in figure 3 for a subset of the cases described in table 1. In agreement with figure 1*E* and *F*, the average mismatch distributions observed in demes with small *Nm* values show two modes, whereas those observed in demes with large *Nm* values ($Nm > 50$) show a single mode. The first mode in demes with low *Nm* values corresponds to the zero-difference class, which is due to pairs of genes with a recent ancestor, whereas the second mode corresponds to pairs of genes having a common ancestor around the time of the onset of the spatial expansion. The 90% empirical confidence intervals for the mismatch distributions presented in figure 3 also show that the variance of the mismatch distributions is much larger for low than for large *Nm* values. In figure 4, we report four random simulated mismatch distributions for demes with either low (10) or large (500) *Nm* values. We see that while the average mismatch distribution for low *Nm* values is bimodal, single realizations of the coalescent in such cases can lead to multimodal and very ragged distributions. In contrast, the mismatch distributions in demes with large *Nm* values are most often unimodal and closer to their expectation, in agreement with the reduced variance shown in figure 3.

In figure 5, we report the mismatch distributions obtained for very different combinations of carrying capacities (*K*) and *m* values leading to the same *Nm* value at equilibrium (when $N = K$). It is clear from this figure that the average shape of the mismatch distributions (and therefore the underlying coalescent process) depends mainly on the value of the product $N \times m$ and almost not on the absolute values of deme size or migration rate. We note, however, that for a given low *Nm* value, there is a slight decrease in the zero-frequency class with larger *N*

Table 1
Summary Statistics Describing the Pattern of Polymorphism Found in a Sample of 30 DNA Sequences

Nm	π[a]	Var(π)	S[b]	Var(S)	D[c]	$P(D) < 0.05$[d]	F_S[e]	$P(F_S) < 0.05$[f]
5	6.3	13.4	30.0	57.0	−0.55	0.03	1.56	0.00
10	7.0	11.7	41.7	75.7	−1.20	0.26	−0.42	0.01
20	7.5	9.8	56.9	85.9	−1.77	0.87	−3.63	0.29
50	7.8	8.3	76.4	97.6	−2.25	1.00	−11.38	0.99
100	7.9	7.8	85.2	88.0	−2.40	1.00	−15.93	1.00
200	8.0	7.6	90.8	72.8	−2.48	1.00	−19.61	1.00
250	8.0	7.5	93.1	74.6	−2.52	1.00	−21.81	1.00
500	8.0	7.5	96.0	69.2	−2.55	1.00	−23.13	1.00
1000	8.0	7.3	96.9	71.4	−2.57	1.00	−23.98	1.00

NOTE.—The sequences are 300 bp drawn from a single deme after a spatial expansion that occurred $\tau = 2Tu = 8$ units of times ago. In this case $T = 4{,}000$ generations, $u = 0.001$, and the sampled deme was located in the center of the array of demes shown in figure 1*A* and *B*, at the same location as the origin of the expansion.

[a] Mean number of differences between all pairs of sequence in the sample.

[b] Number of segregating sites.

[c] Tajima's D statistic (Tajima 1989).

[d] Probability that Tajima's D statistic is found significant at the 5% level estimated from 1,000 simulations.

[e] Fu's F_S statistic (Fu 1997).

[f] Probability that Fu's F_S statistic is found significant at the 5% level estimated from 1,000 simulations.

values (fig. 5, left column with $K = 500$ and K=1,000, as compared with K=100). Note that no such effect is observed for large Nm values, as shown in the right column of figure 5. This phenomenon may be due to the fact that with low N values (implying a large m value), several gene lineages may initially comigrate in the same deme and subsequently coalesce, whereas with smaller m values, gene lineages will migrate once at a time. The comigration of genes in the same deme thus slightly increases the probability of recent coalescent events, leading to the slightly larger probabilities of no differences between genes sampled in demes of small size and exchanging a large fraction of genes with their few neighbors.

The age of the expansion seems to affect the pattern of diversity in a more drastic way for low than for large Nm values (table 3). For $Nm = 500$, Tajima's D and Fu's F_S statistics are very efficient in detecting departure from population equilibrium, irrespective of the age of the expansion. In contrast, for $Nm = 10$, Tajima's D statistic is much less powerful, showing departure from equilibrium between only one fourth and one third of the cases. For the same amount of gene flow, the behavior of the test based on Fu's F_S statistic is markedly different. The hypothesis of selective neutrality and population equilibrium will be more often rejected for relatively recent expansions ($\tau < 3$) than for older expansions ($\tau > 5$), which is somewhat counterintuitive. However, we may propose the following explanation. Since Fu's F_S statistic is the logit of the probability to observe k or more alleles conditional on π, the observed average number of pairwise differences, the behavior of this test can be explained by understanding the behavior of k and π under spatial expansions. As visible on the first row of figure 2, a range expansion with limited gene flow among demes produces an intra-deme gene genealogy with both many recent and many old coalescent events. The age of the old coalescent events depend essentially of the age of the expansion, whereas the age of the recent coalescent events depends on the size of the deme. For a sufficiently large mutation rate, the age of the expansion will not affect much k, but it will have a large effect on π. Thus, the probability of observing a given number k or more alleles will increase with older expansions, leading to less negative F_S values, as shown in table 3. The effect of the age of the expansion on the mismatch distribution is clearer, and much like in the case

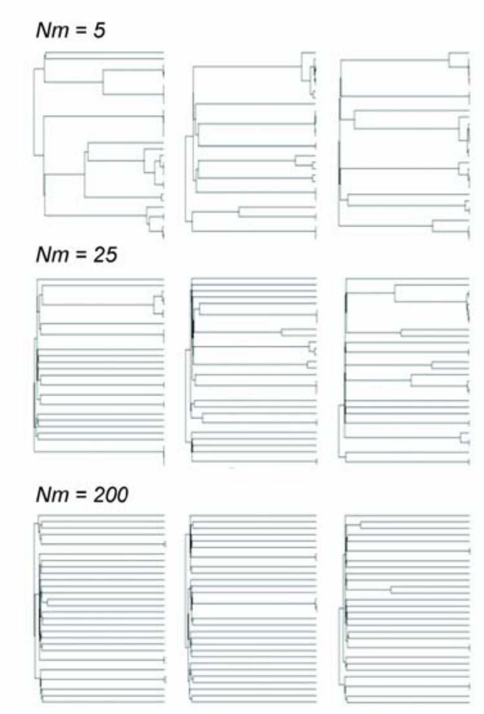


FIG. 2.—Gene genealogies after a spatial expansion. Three random genealogies of 30 genes are shown for Nm values of 5, 25, and 200 migrants exchanged between neighboring demes. The spatial expansion occurred $\tau = 8$ units of time ago, as indicated in the footnote of table 1.

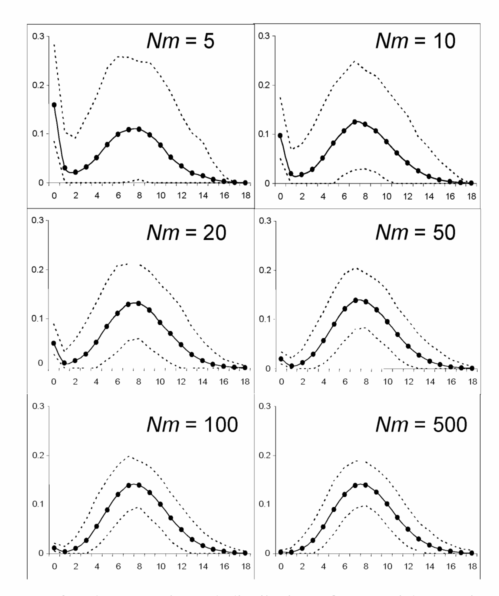


FIG. 3.—Average mismatch distributions after a spatial expansion for different *Nm* values. The Y axis stands for the average probability that two DNA sequences differ at a given number of sites represented on the X axis. The solid lines are average mismatch distributions obtained from 1,000 simulations of the coalescent of 30 genes drawn in a single deme after a spatial expansion having occurred $\tau = 8$ units of time ago. Dotted lines delimit an empirical 90% confidence interval for the mismatch distribution.

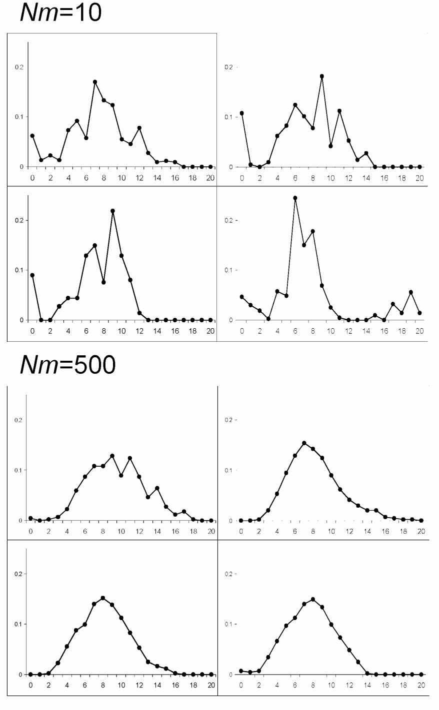


FIG. 4.—Mismatch distributions obtained for single realizations of the coalescent after a spatial expansion for demes exchanging either a low (*Nm*=10) or a large (*Nm* = 500) numbers of genes each generation with neighboring demes.

of demographic expansion in unsubdivided populations (Rogers and Harpending 1992), the mismatch distribution mode is shifted to the right with older expansion times (data not shown).

Discussion

Implication for Human Mitochondrial DNA Diversity

Previous interpretations concerning the pattern of diversity in mitochondrial mtDNA have relied on the assumption that populations were unsubdivided (Slatkin and Hudson 1991; Rogers and Harpending 1992; Rogers 1995; Weiss, Henking, and von Haeseler 1997; Excoffier and Schneider 1999), with some exceptions (e.g., Marjoram and Donnelly 1994). Under this paradigm, unimodal mismatch distributions have been interpreted as being due to past demographic expansions (Slatkin and Hudson 1991; Rogers and Harpending 1992). However, although it is true that most human populations show approximately unimodal mismatch distributions compatible with Pleistocene population expansions (fig. 6*A*), almost all present or recent hunter-gatherer groups show very ragged distributions and in particular a high proportion of pairs of sequences that are similar, thus showing no differences (fig. 6*B*). This contrast has been interpreted as the consequence of a recent (post-Neolithic) contraction

Table 2
Influence of the Sampling Location and the Expansion Origin on Patterns of Molecular Diversity

Expansion Origin	Sampling Location	*Nm*	π	Var(π)	*S*	Var(*S*)	*D*	$P(D) < 0.05$	F_{Sz}	$P(F_S) < 0.05$
Periphery . .	Periphery	10	6.6	11.8	34.7	60.5	−0.86	0.08	−0.04	0.00
Periphery . .	Center	10	6.9	11.0	40.3	71.7	−1.16	0.22	−0.52	0.01
Center	Periphery	10	6.6	11.8	34.3	54.6	−0.82	0.07	−0.09	0.01
Center	Center	10	7.0	11.7	41.7	75.7	−1.20	0.26	−0.42	0.01
Periphery . .	Periphery	500	8.0	7.4	94.4	65.0	−2.53	1.00	−22.89	1.00
Periphery . .	Center	500	8.0	7.4	95.7	65.9	−2.55	1.00	−23.21	1.00
Center	Periphery	500	7.9	7.3	93.2	77.6	−2.53	1.00	−23.10	1.00
Center	Center	500	8.0	7.5	96.0	69.2	−2.55	1.00	−23.13	1.00

NOTE.—Center refers to the central deme in our simulated array of 50 × 50 demes. It is thus located at position <25; 25>. Periphery refers to a deme located in the periphery of the simulated array, at position <5; 5>. The remaining headers are identical to those described in table 1.

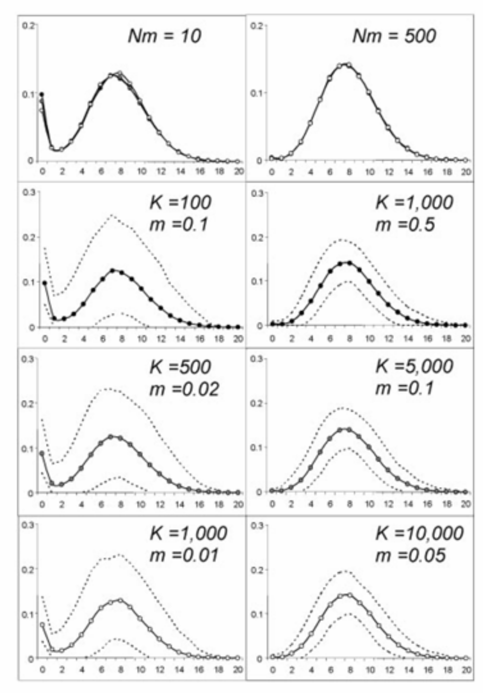
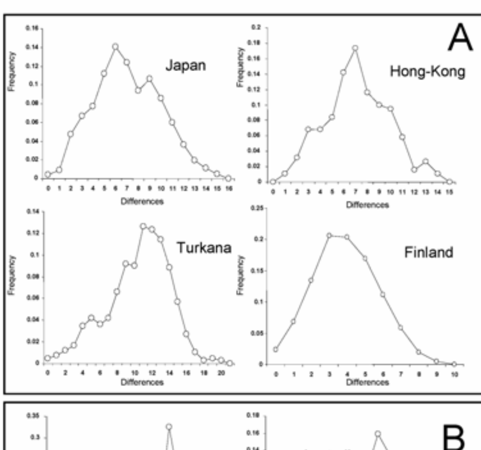


FIG. 5.—Mismatch distributions obtained for three different combinations of carrying capacities (*K*) and proportion of migrants exchanged with neighboring demes (*m*), leading to similar *Nm* values at carrying capacity ($N = K$). Left panels: $Nm = 10$; right panels: $Nm = 500$. The averaged mismatch distributions corresponding to the three different cases are superimposed on the top panels and are shown separately with their 90% confidence intervals on the three lower panels.



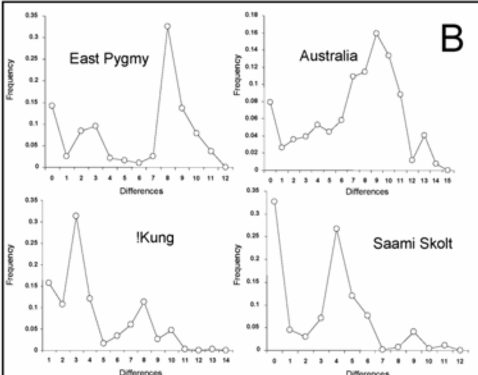


FIG. 6.—Observed mismatch distributions in human populations analyzed for mtDNA hypervariable region 1 (HVR1). Data drawn from samples referenced in Excoffier and Schneider (1999). *A*, Post-Neolithic populations. *B*, Present or former hunter-gatherer populations.

of the size of hunter-gatherer populations, resulting from the fragmentation of their habitat leading to contraction of their effective size (Excoffier and Schneider 1999).

Our present results would however lead to a simpler and very different interpretation of the differences in the shape of mismatch distribution between post-Neolithic and hunter-gatherer populations. By assuming that the present distribution of human populations results from some spatial range expansion, this contrast would simply result from the much larger deme size of Neolithic populations (resulting in much larger *Nm* values) than hunter-gatherer populations. While our simulations have assumed constant deme sizes from the onset of the range expansion to the present time, it is easy to simulate a range expansion with

Table 3
Different Statistics Summarizing the Pattern of Molecular Diversity After Range Expansion

Nm	Range Expansion[a]	π	Var(π)	*S*	Var(*S*)	*D*	$P(D) < 0.05$	F_S	$P(F_S) < 0.05$
10	1	1.1	0.9	7.1	8.0	−1.17	0.31	−2.74	0.57
	3	2.8	3.0	17.9	23.7	−1.28	0.32	−3.38	0.45
	5	4.5	5.8	27.6	40.3	−1.25	0.29	−2.21	0.17
	7	6.1	9.5	36.7	62.9	−1.19	0.26	−0.90	0.02
	8	7.0	11.7	41.7	75.7	−1.20	0.26	−0.42	0.01
500	1	1.2	1.1	14.5	13.6	−2.26	0.99	−11.50	1.00
	3	3.1	2.9	40.4	37.2	−2.55	1.00	−23.06	1.00
	5	5.1	4.8	64.4	48.0	−2.58	1.00	−24.55	1.00
	7	7.1	6.8	85.9	75.0	−2.56	1.00	−24.06	1.00
	8	8.0	7.5	96.0	69.2	−2.55	1.00	−23.13	1.00

NOTE.—The different times of expansions are $T = 500$, 1,500, 2,500, 3,500, and 4,000 generations ago, and the different migration intensities between neighboring demes are $Nm = 10$ and 500.

[a] Date of the onset of the range expansion τ, in units of mutation rate *u*, as $\tau = 2Tu$, where *T* is the time of the expansion in number of generations, and $u = 0.001$.

Table 4
Pattern of Molecular Diversity After a Spatial Expansion

#	K_0	K_1	Size Exp.	π	Var(π)	S	Var(S)	D	$P(D) < 0.05$	F_S	$P(F_S) < 0.05$
A	1000	1000	—	9.6	11.7	90.1	95.6	−2.20	1.00	−13.9	1.00
B	100	1000	500	7.8	8.1	73.4	81.2	−2.19	1.00	−14.2	1.00
C	100	1000	100	7.4	8.8	60.4	85.0	−1.91	0.95	−8.1	0.85
D	100	1000	50	7.5	9.6	56.0	84.8	−1.75	0.84	−5.3	0.54
E	100	1000	10	7.1	11.4	43.3	74.1	−1.26	0.29	−0.8	0.01
F.	100	100	—	7.0	11.7	41.7	75.7	−1.20	0.26	−0.4	0.01

NOTE.—The expansion started 4,000 generations ago and was followed by a more recent global demographic expansion at different times in the past (T = 10, 50, 100, and 500 generations ago). A fraction m = 0.10 of migrants are constantly exchanged between neighboring demes. K_0 = Carrying capacity of the demes before the demographic expansion. K_1 = Carrying capacity of the demes after the demographic expansion. Size Exp. is the time in generations (before present) at which the demographic expansion occurs.

small deme size and a recent increase in deme size resulting in higher levels of gene flow with surrounding demes. The results of such simulations are shown in table 4 for the pattern of molecular diversity and in figure 7 for the mismatch distributions.

We find that demographic expansions having occurred more than 100 generations ago and resulting in a 10-fold increase in Nm values (from $Nm = 10$ to $Nm = 100$) would lead to unimodal distributions (fig. 7*B* and *C*), as if their size had always been 10-fold higher (fig. 7*A*). In contrast, more recent demographic expansions would lead to a greater number of recent coalescent events and multimodal distributions (fig. 7*D* and *E*), as if deme size had always been low (fig. 7*F*). Patterns of molecular diversity show a very similar trend (table 4), with demographic expansions having occurred more than 50 generations ago resulting in a clear rejection of neutrality and population equilibrium with Tajima's D or Fu's F_S statistic. These simulations clearly show that relatively recent demographic expansions leading to overall larger Nm values lead to patterns of molecular diversity equivalent to those expected in demes having always exchanged large numbers of individuals with their neighbors. It thus seems that the Nm value prevailing during the scattering phase (*sensu* Wakeley 1999) of the gene lineages is the factor that will primarily determine the overall pattern of diversity observed within demes. Large Nm values during the scattering phase are sufficient to prevent recent coalescent events. The ancestral lineages of almost all sampled genes will thus be found in different demes at the end of the scattering phase. After that point, if the number of demes is much larger than the number of remaining lineages, the size of the demes (and their associated Nm value) will have almost no effect on the pattern of coalescence until the onset of the spatial expansion. This property should make the model quite robust to the likely complex histories of natural populations going through long-term size fluctuations.

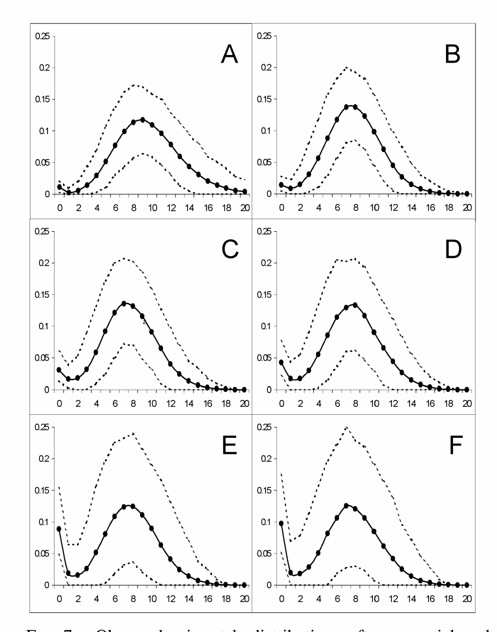


FIG. 7.—Observed mismatch distributions after a spatial and an optional demographic expansion. Cases *A* through *F* correspond to demographic histories defined in table 4. *A*, Demes with constant K = 1,000. *B*, Demographic growth from K_0 = 100 to K_1 = 1,000 occurring 500 generation before present. *C*, Same as *B* but demographic growth occurring 100 generations before present. *D*, Same as *B* but demographic growth occurring 50 generations before present. *E*, Same as *B* but demographic growth occurring 10 generations before present. *F*, Demes with constant K = 100.

The present simulation results thus explain the difference between the mismatch distributions of hunter-gatherer and post-Neolithic populations by the simple fact that food gatherers have generally lower densities than food producers (if one assumes that both groups have approximately similar emigration rates). However, additional factors may have led to different patterns of molecular diversity in these communities. It remains true that present hunter-gatherer communities currently live in environments that are unfavorable and more fragmented than before (Lewin 1988), which could have reduced considerably their effective population size and thus led to multimodal mismatch distribution (Excoffier and Schneider 1999). Such a process would certainly reinforce the difference in recent deme size between the two types of communities and contribute to the extreme raggedness of hunter-gatherer mismatch distributions. But we feel that a realistic model of population differentiation should necessarily take into account the subdivision of human populations. Therefore, a scenario with global demographic growth and subsequent bottlenecks to explain observed differences between patterns of diversity in food-producing and food-gathering populations appears much less par-

simonious and less likely than simply taking into account the finite spatial structure of the demes and the low census size of hunter-gatherers.

Distinction Between Spatial and Demographic Expansions

We find that although spatial expansions also involve a demographic increase at the level of the population as a whole, they do not necessarily lead to a molecular signature similar to that of sudden demographic expansions in unsubdivided populations. This is the case only if the amount of gene flow is large between neighboring demes. For relatively low levels of gene flow ($Nm < 20$), recent coalescent events and therefore multimodal mismatch distributions can be expected in a quite large fraction of simulations (table 1), even if the global size of the population has been increased by several orders of magnitude after the expansion. The dependence between the amount of gene flow between demes and the average level of genetic diversity (π) within deme observed after a spatial expansion is different from that expected in a subdivided population at equilibrium. Several studies have indeed shown that the average coalescence time between a pair of genes should only depend on the total size of the population, if demes are all either directly or indirectly interconnected (Slatkin 1987; Strobeck 1987; Hey 1991) and if the number of demes is constant (Nagylaki 1998). Examination of table 1 suggests that demes with low levels of gene flow should show lower average levels of diversity (both lower π and lower S values) than demes with high gene flow after a spatial or range expansion. Also note that what we call "low levels of gene flow" are still cases where Nm is much greater than 1, which is generally the value above which spatially arranged demes are assumed to evolve as a single unit (e.g., Maruyama 1971). This result underlines the need to further study spatial models of populations out of equilibrium.

Another prediction that may reveal differences between models of demographic and spatial expansions is the relationship between the geographical location of the sample and its genetic diversity. Results shown in table 2 suggest that demes sampled in the periphery of the present population range may show slightly reduced levels of molecular diversity for low Nm values, regardless of the origin of the expansion. This may be due to the fact that gene lineages are less free to diffuse to different demes in the scattering phase when they are close to the border of the expansion range. They would thus spend more time within the same deme and have therefore more time to coalesce. The spatial diffusion constraints during the scattering phase would lead to an excess of recent coalescent events as compared with genes sampled in more central demes. This suggest that the pattern of molecular diversity within samples should be affected by the presence of geographical barriers preventing a free diffusion of genes to neighboring demes for species having low dispersal abilities. Note that this effect would be quite different from the reduced diversity expected in marginal populations and resulting from a demic diffusion process from a given source (Rendine, Piazza, and Cavalli-Sforza 1986; Sokal, Oden, and Wilson 1991; Barbujani, Sokal, and Oden 1995), where one would expect a loss of genetic diversity due to a succession of small founder effects. However, a clearer distinction between demographic and spatial expansions should emerge from the study of samples of genes taken from different demes, which should be the object of a different study.

Recent Range Expansions as a Way to Examine Patterns of Dispersal from Single Samples

Recent range expansions and speciations are thought to have been quite common in the Quaternary, following or due to ice ages, respectively (for a review, see e.g., Taberlet et al. 1998; Hewitt 2000). It is therefore likely that the traces of recent spatial expansions could be found in many species other than humans, in fact in all populations that would have gone through very small sizes during former ice ages spent in refuge areas, from where they would have then reexpanded. Interestingly, the fact that some populations would have expanded from a refuge area would not only tell us something about their global dispersal abilities but could also bring important information on their recent rate of dispersal outside their demes. Since the shape of the mismatch distribution, and particularly the frequency of recent coalescent events, depends on recent migration rates, it should be possible to estimate emigration rates by sampling individuals from the same deme and examining their pattern of molecular diversity. Applied to sex-linked markers, this could allow one to study potential sex-biased dispersal and/or different effective size between sexes. An estimation procedure for Nm values inferred from a single sample drawn from a recently expanding population is currently under investigation, and it will be the subject of a forthcoming paper. Available methods for estimating levels of gene flow usually rely on the availability of a series of samples. Gene flow is then inferred between demes from which the samples are supposed to be drawn (see e.g., Beerli and Felsenstein 2001). This implies that sampled demes actually exchange migrants and that one is able to define the geographical limit of the deme. The validity of these two assumptions is generally quite difficult to assess and would not be required from the analysis of single samples. We are therefore confident that the analysis of patterns of molecular diversity from single deme samples would allow one to get important insights on the life history of the numerous populations having gone through recent range expansions.

Acknowledgments

Thanks to Pierre Berthier and Stefan Schneider for their computing and programming assistance. We are grateful to Grant Hamilton for his careful reading of the manuscript and to Montgomery Slatkin and Henry Harpending for their comments on an earlier version of the manuscript. We are also indebted to John Wakeley for his many suggestions helping to improve various aspects of the manuscript. This work was supported by a Swiss NSF grant No 31–054059–98 to L.E.

Literature Cited

Aris-Brosou, S., and L. Excoffier. 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. **13**: 494–504.

Barbujani, G., R. R. Sokal, and N. L. Oden. 1995. Indo-European origins: a computer-simulation test of five hypotheses. Am. J. Physical Anthropol. **96**:109–132.

Beaumont, M. A. 1999. Detecting population expansion and decline using microsatellites. Genetics **153**:2013–2029.

Beerli, P., and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152**:763–773.

———. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA **98**:4563–4568.

Donnelly, P., and S. Tavaré. 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29**:401–421.

Excoffier, L., J. Novembre, and S. Schneider. 2000. SIMCOAL: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J. Hered. **91**:506–510.

Excoffier, L., and L. Schneider. 2000. The demography of human populations inferred from patterns of mitochondrial DNA diversity. Pp. 101–108 *in* C. Renfrew and K. Boyle, eds. Archaeogenetics: DNA and the population prehistory of Europe. McDonald Institute for Archeological Research, Cambridge.

Excoffier, L., and S. Schneider. 1999. Why hunter-gatherer populations do not show sign of Pleistocene demographic expansions. Proc. Natl. Acad. Sci. USA **96**: 10597–10602.

Fu, Y.-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. Genetics **147**:915–925.

Goldstein, D. B., G. W. Roemer, D. A. Smith, D. E. Reich, A. Bergman, and R. K. Wayne. 1999. The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics **151**: 797–801.

Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers, and S. T. Sherry. 1998. Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95**:1961–1967.

Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. Nature **405**:907–13.

Hey, J. 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. Theor. Popul. Biol. **39**:30–48.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. J. Futuyma and J. D. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, New York.

Kingman, J. F. C. 1982*a*. The coalescent. Stoch. Proc. Appl. **13**: 235–248.

———. 1982*b*. On the genealogy of large populations. J. Appl. Proba. **19A**:27–43.

Lewin, R. 1988. New views emerge on hunters and gatherers. Science **240**:1146–1148.

Lundstrom, R., S. Tavaré, and R. H. Ward. 1992. Modeling the evolution of the human mitochondrial genome. Math. Biosci. **112**:319–335.

Marjoram, P., and P. Donnelly. 1994. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136**: 673–683.

Maruyama, T. 1971. Analysis of population structure. II. Two-dimensional stepping stone models of finite lengths and other geographically structured populations. Ann. Hum. Genet. **35**:179–196.

Nagylaki, T. 1998. The expected number of heterozygous sites in a subdivided population. Genetics **149**:1599–1604.

Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154**:931–942.

Nordborg, M. 1997. Structured coalescent processes on different time scales. Genetics **146**:1501–1514.

———. 2001. Coalescent theory. Pp. 179–212 *in* D. Balding, M. Bishop, and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, New York.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population. J. Math. Biol. **29**: 59–75.

Pereira, L., I. Dupanloup, Z. H. Rosser, M. A. Jobling, and G. Barbujani. 2001. Y-chromosome mismatch distributions in Europe. Mol. Biol. Evol. **18**:1259–1271.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16**:1791–1798.

Reich, D. E., and D. B. Goldstein. 1998. Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95**:8119–8123.

Rendine, S., A. Piazza, and L. L. Cavalli-Sforza. 1986. Simulation and separation by principal components of multiple demic expansions in Europe. Am. Nat. **128**:681–706.

Rogers, A. 1995. Genetic evidence for a Pleistocene population explosion. Evolution **49**:608–615.

Rogers, A. R., and H. Harpending. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9**:552–569.

Rogers, A. R., and L. B. Jorde. 1995. Genetic evidence on modern human origins. Hum. Biol. **67**:1–36.

Rousset, F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142**:1357–1362.

———. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics **145**:1219–1228.

Schneider, S., and L. Excoffier. 1999. Estimation of demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152**:1079–1089.

Schneider, S., D. Roessli, and L. Excoffier. 2000. ARLEQUIN: a software for population genetics data analysis. Version 2.000. University of Geneva, Geneva, Switzerland.

Sherry, S. T., A. R. Rogers, H. Harpending, H. Soodyall, T. Jenkins, and M. Stoneking. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. Hum. Biol. **66**:761–775.

Slatkin, M. 1987. The average number of sites separating DNA sequences drawn from a subdivided population. Theor. Popul. Biol. **32**:42–49.

———. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

Slatkin, M., and R. R. Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**:555–562.

Sokal, R. R., N. L. Oden, and C. Wilson. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature **351**:143–145.

Strobeck, K. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117**:149–153.

Taberlet, P., L. Fumagalli, A. G. Wust-Saucy, and J. F. Cosson. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Mol. Ecol. **7**:453–464.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585–595.

Wakeley, J. 1999. Nonequilibrium migration in human history. Genetics **153**:1863–1871.

———. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution **54**:1092–1101.

———. 2001. The coalescent in an island model of population subdivision with variation among demes. Theor. Popul. Biol. **59**:133–144.

Wakeley, J., and N. Aliacar. 2001. Gene genealogies in a metapopulation. Genetics **159**:893–905.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero, and K. Ardlie. 2001. The discovery of single-nucleotide polymorphisms-and inferences about human demographic history. Am. J. Hum. Genet. **69**:1332–1347.

Watson, E., K. Bauer, R. Aman, G. Weiss, A. von Haeseler, and S. Paabo. 1996. mtDNA sequence diversity in Africa. Am. J. Hum. Genet. **59**:437–444.

Weiss, G., A. Henking, and A. von Haeseler. 1997. Distribution of pairwise differences in growing populations. Pp. 81–95 *in* P. Donnely and S. Tavaré, eds. Progress in population genetics and human evolution. Springer Verlag, New York.

Wilkinson-Herbots, H. M. 1998. Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. **37**:535–585.

Naruya Saitou, Associate Editor

Accepted September 13, 2002

3.3 Signature d'une expansion spatiale dans les données moléculaires de type SNP

Comme nous l'avons vu dans la section précédente (3.2), la distribution "mismatch" présente une forme différente lorsqu'elle est tirée d'une population ayant passé par une croissance démographique et lorsqu'elle est tirée d'une population stationnaire (Rogers et Harpending 1992 ; Harpending *et al.* 1993 ; Harpending 1994). Cette statistique a été abondamment utilisée pour étudier le génome mitochondrial des populations humaines (Di Rienzo et Wilson 1991 ; Bertranpetit *et al.* 1995 ; Sajantila *et al.* 1995 ; Calafell *et al.* 1996 ; Comas *et al.* 1996 ; Corte-Real *et al.* 1996 ; Francalacci *et al.* 1996 ; Comas *et al.* 2000 ; Malyarchuk et Derenko 2001 ; Nasidze et Stoneking 2001), puisque celles-ci sont principalement composées de séquences d'ADN pour lesquelles cette approche a été développée. Le génome mitochondrial étant transmis uniquement par voie maternelle, il ne donne des informations que sur la démographie de la lignée féminine. Il est donc tentant de comparer les distributions "mismatch" obtenues pour l'ADN mitochondrial avec celles obtenues pour la lignée masculine. Malheureusement, les séquences complètes d'ADN pour le système génétique spécifique aux mâles – la partie non-recombinante du chromosome Y, ou MSY – sont rares (Whitfield *et al.* 1995 ; Shen *et al.* 2000 ; Hammer *et al.* 2003). Les données génétiques disponibles pour ce système sont principalement composées de microsatellites (de Knijff *et al.* 1997 ; Pritchard *et al.* 1999 ; Belledi *et al.* 2000 ; Forster *et al.* 2000 ; Shen *et al.* 2000 ; Kayser *et al.* 2001), de SNPs (Semino *et al.* 2000a ; Bosch *et al.* 2001 ; Hammer *et al.* 2001 ; Underhill *et al.* 2001 ; Shen *et al.* 2002), ou des deux (Bosch *et al.* 1999 ; Carvajal-Carmona *et al.* 2000 ; Al-Zahery *et al.* 2003). Des distributions "mismatch" pour les populations européennes et moyen-orientales ont tout de même été établies sur la base de SNPs (Pereira *et al.* 2001 ; Dupanloup *et al.* 2003). Ces derniers sont sujets à un important "biais de recrutement" ("ascertainment bias" en anglais, Rogers et Jorde 1996) car ils ne correspondent pas à tous les sites polymorphes qui existent dans les séquences d'ADN. En effet, les mutations les plus rares dans la population sont sous-représentées dans les échantillons constitués de SNPs (voir page 182). Il nous a donc paru intéressant d'utiliser SPLATCHE pour étudier l'effet de l'utilisation de SNPs dans la constitution de distributions "mismatch", ainsi que d'étudier l'effet du biais de recrutement sur ces distributions. Dans un deuxième temps, nous discuterons ces résultats, en fonction des données connues pour les populations européennes.

3.3.1 Simulations de séries de SNPs pour différents types d'expansion

Nous avons procédé à une série de simulations selon le même schéma que celui décrit dans la section précédente (3.2), soit dans un monde virtuel carré (50x50), homogène pour les facteurs environnementaux (*K* et *F*). L'expansion spatiale d'une population – dont le *Nm* est soit petit (*Nm* = 2), soit grand (*Nm* = 100) – est simulée depuis le dème central <25; 25> pendant 4'000 générations. Ces deux types de populations représentent, respectivement, une population de chasseurs-collecteurs actuels (petit *Nm*) et une population post-néolithique (grand *Nm*). Différents effectifs de SNPs (10, 50 et 100) sont simulés et leur distribution "mismatch" établie. Ci-dessous, nous

présentons les résultats divisés en deux catégories : 1° résultats pour lesquels tous les SNPs générés sont utilisés (aucun biais de recrutement) ; 2° résultats pour lesquels seuls des sites dont la fréquence allélique est supérieure ou égale à 10% sont étudiés, afin de tenir compte du biais de recrutement (voir page 182).

Nos simulations montrent qu'il est possible d'obtenir des informations sur la démographie des populations en utilisant uniquement les sites polymorphes d'une séquence, lorsqu'aucun choix de SNPs n'est fait. Les distributions "mismatch" possèdent alors les mêmes caractéristiques que lorsque des séquences entières sont utilisées: *i)* bimodale accompagnée d'une grande variance lorsque *Nm* est petit (Figure 3.1 a, c et e) ; *ii)* unimodale accompagnée d'une faible variance lorsque *Nm* est grand (Figure 3.2 a, c et e). La grande différence entre les distributions "mismatch" tirées des séquences complètes et celles obtenues sur la base de SNPs est que ces dernières ne livrent aucune indication quant à la date de l'expansion puisque les SNPs sont indépendants du taux de mutation. En effet, plus le nombre de SNPs utilisés est important et plus le mode "principal" de la distribution "mismatch" est important. Nous définissons le mode "principal" de la distribution "mismatch" comme celui qui est généré par les coalescences qui ont lieu au moment de l'origine de l'expansion à l'intérieur ou autour du dème ancestral ("phase de contraction"). Par opposition, le "premier" mode est généré par les coalescences récentes qui ont lieu dans le dème d'échantillonnage. Ce "premier" mode se traduit par une classe 0 importante et correspond à l'homozygotie[1] attendue de la population.

La valeur du mode "principal" des distributions "mismatch" tirées de SNPs est beaucoup plus élevée lorsque *Nm* est petit (Figure 3.1) que lorsqu'il est grand (Figure 3.2). Cette différence s'explique par le fait que le nombre de SNPs, donc de sites polymorphes *S*, est identique dans les deux cas, mais que la forme de la généalogie est différente (Figure 3.3). Lorsque *Nm* est grand, les mutations se répartissent principalement sur les branches terminales qui sont très longues. Il y a par conséquent très peu de gènes identiques mais leur différenciation est modérément importante lorsqu'ils sont comparés deux à deux. A l'opposé, la longueur importante des branches internes dans les généalogies obtenues avec un petit *Nm* implique que la majorité des mutations s'accumulent sur ces branches internes. Il en résulte que les gènes pris par paires sont soit identiques, soit très différents. Lorsque *Nm* est petit, le mode "principal" de la "mismatch" reflète la grande différenciation qui existe entre une partie des gènes; sa valeur augmente avec la diminution de *Nm*.

Cette variation du mode "principal" des distributions "mismatch" en fonction du *Nm* ne s'observe pas avec des séquence d'ADN (article 3.2.1 : Figure 3), car avec ces dernières le nombre de sites polymorphes *S* n'est pas fixé. La taille totale de l'arbre de coalescence augmente avec *Nm* et par conséquent *S* également (article 3.2.1 : Table 1). Pour la même longueur de séquence, la présence

---

[1] Il ne s'agit pas ici d'homozygotie réelle puisque le locus simulé est haploïde mais nous utilisons cependant ce terme pour décrire la proportion de séquences identiques attendue dans la population.

d'un plus grand nombre de sites polymorphes *S* lorsque *Nm* est grand, compense la valeur plus élevée du mode principal lorsque *Nm* est petit.

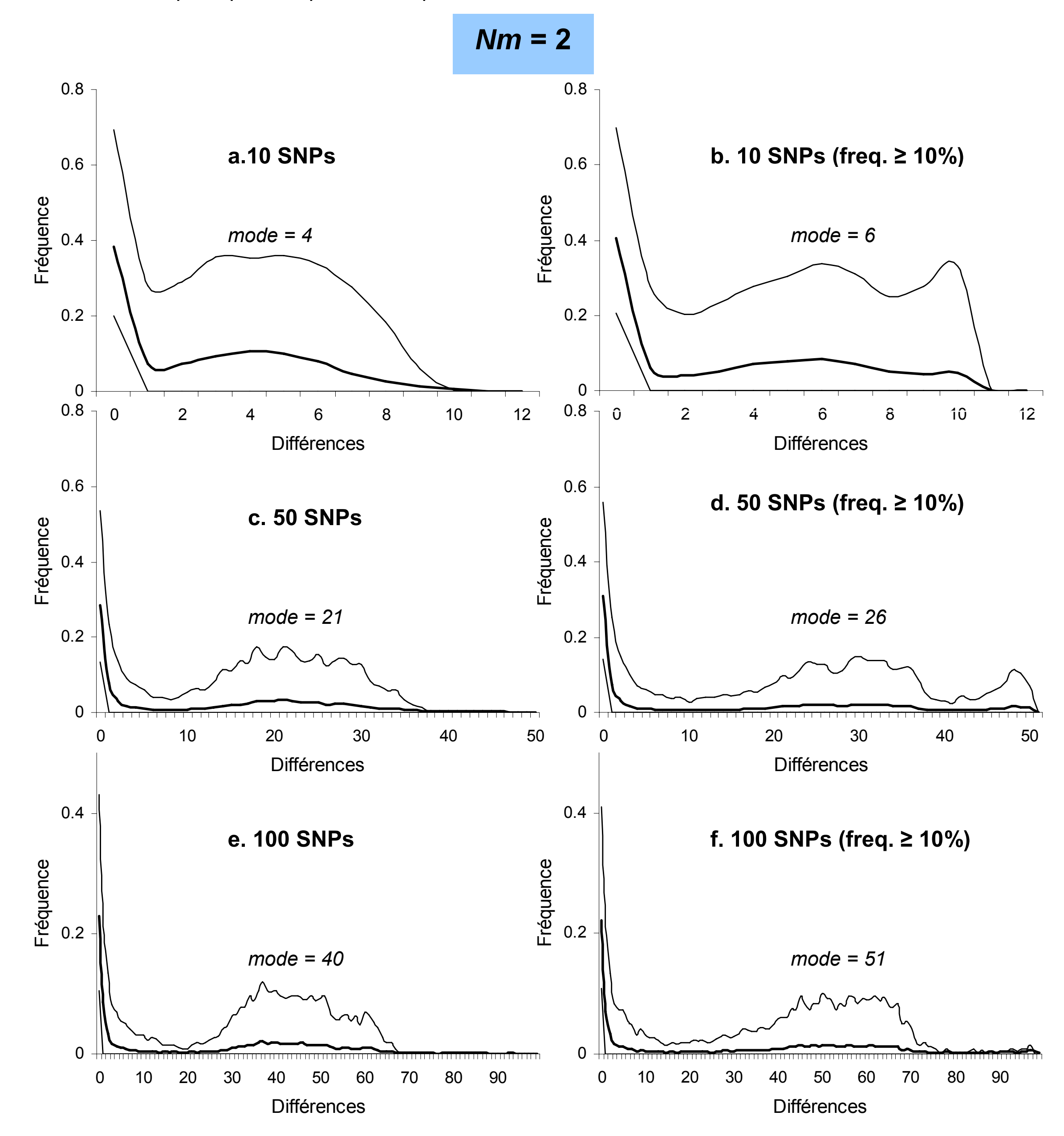


Figure 3.1 Distribution "mismatch" moyenne (ligne épaisse) pour 1'000 simulations, et intervalle de confiance à 90% (lignes fines). Colonne de gauche : tous les SNPs sont considérés. Colonne de droite : SNPs dont la fréquence de l'allèle le plus rare est supérieure ou égale à 10%. La valeur du mode "principal" (voir texte) est donné pour chaque figure.

Le biais de recrutement correspond à la sous-représentation des SNPs dont la fréquence est faible dans la population étudiée (voir page 182). Ce choix revient à sélectionner des mutations qui se trouvent sur les branches internes de la généalogie des gènes. Lorsque *Nm* est petit, alors le biais de recrutement ne change quasiment rien aux distributions "mismatch", si ce n'est que la valeur du mode "principal" est légèrement plus élevée (Figure 3.1 b, d et f). Ceci s'explique par le fait que lorsque *Nm* est faible, les branches internes de l'arbre de coalescence sont longues, les

mutations ont donc une plus grande probabilité de s'y accumuler. Par conséquent, la sélection des SNPs avec les plus hautes fréquences alléliques sera alors plus ou moins représentative de la configuration de la généalogie (Figure 3.3a). En revanche, lorsque *Nm* est grand, le biais de recrutement provoque un "premier" mode important dans les distributions "mismatch" (Figure 3.2 b, d et f), qui résulte de l'utilisation des mutations qui se trouvent sur les branches internes de l'arbre. En effet, lorsque *Nm* est grand la majorité des mutations s'accumulent sur les branches terminales de l'arbre. Par conséquent, l'étude des mutations les plus fréquentes n'est pas représentatif de la configuration de la généalogie (Figure 3.3b).

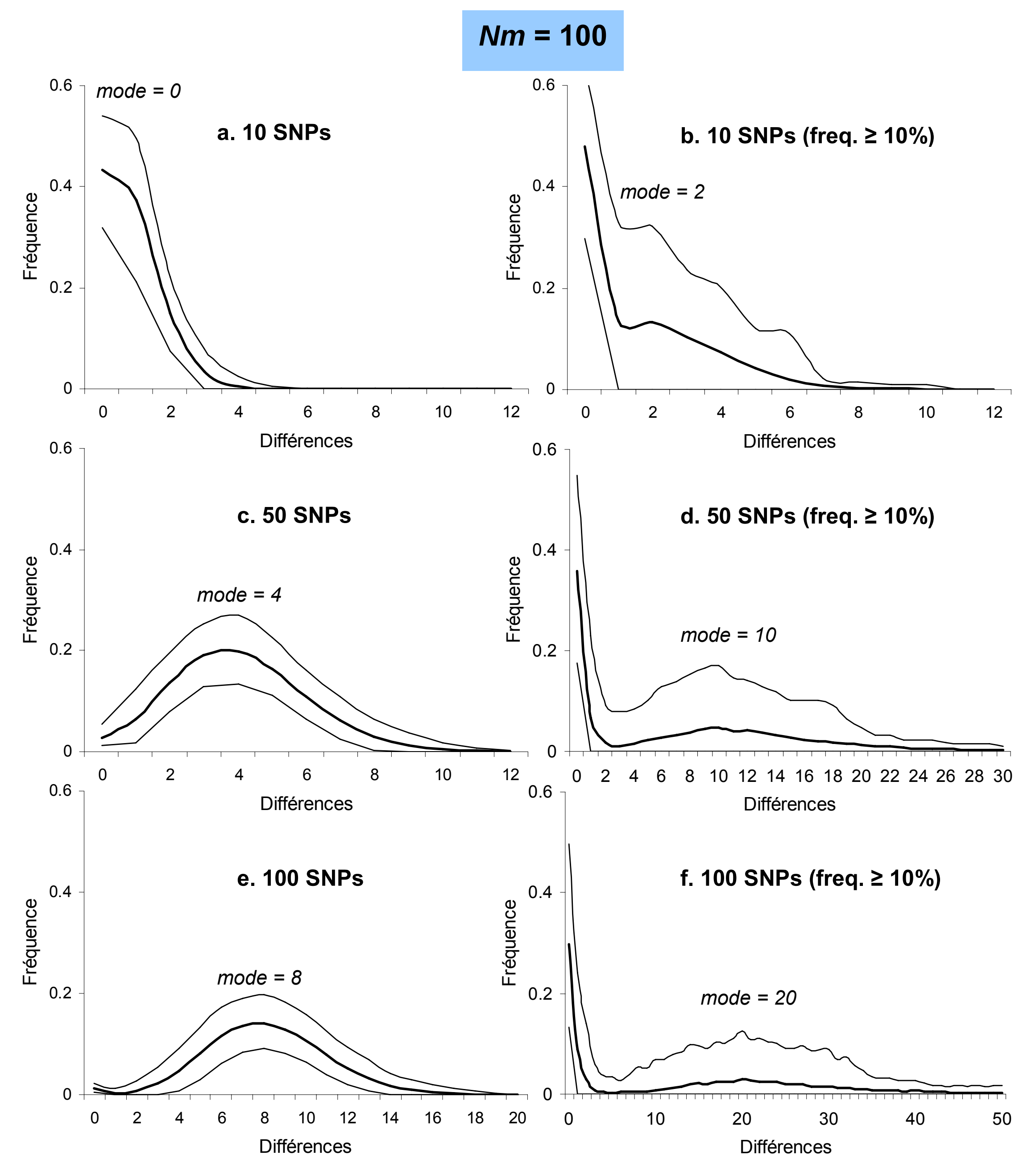


Figure 3.2 Distribution "mismatch" moyenne (ligne épaisse) pour 1'000 simulations et intervalle de confiance à 90% (lignes fines). Colonne de gauche : tous les SNPs sont considérés. Colonne de droite : SNPs dont la fréquence de l'allèle le plus rare est supérieure ou égale à 10%. La valeur du mode "principal" (voir texte) est donnée pour chaque figure.

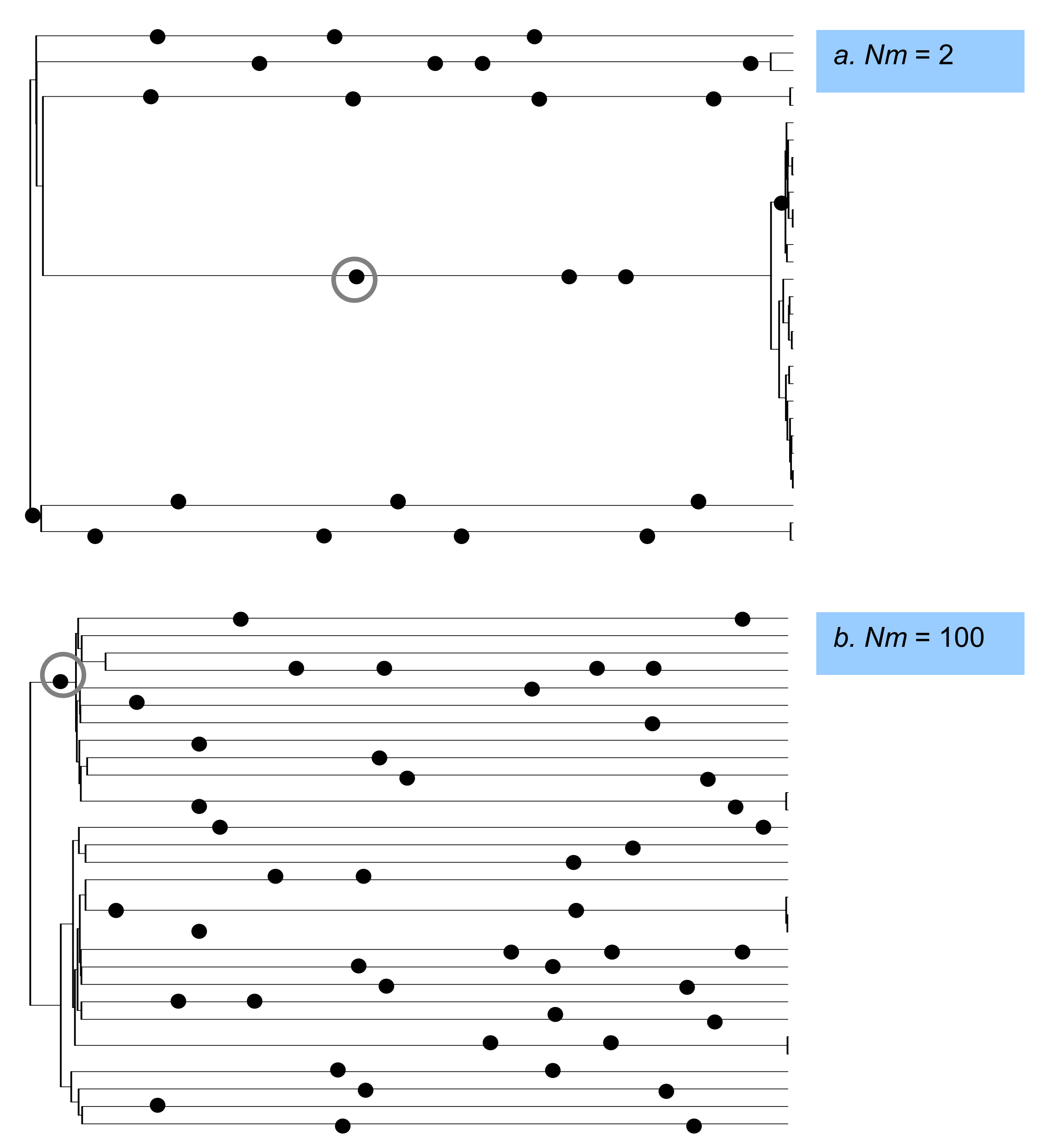


Figure 3.3 Exemples de généalogies de gènes échantillonnés. A : dans le cas d'un petit *Nm* (=2). B : dans le cas d'un grand *Nm* (=100). Le choix des SNPs dont la fréquence est élevée dans la population (cercles gris) porte sur les mutations qui ont lieu sur les branches internes. Il est représentatif des mutations (points noirs) qui s'accumulent également sur ces branches lorsque *Nm* est petit, mais pas lorsqu'il est grand.

Nous avons donc montré ici que les distributions "mismatch" tirées de SNPs permettent d'estimer le *Nm* de la population échantillonnée seulement si le biais de recrutement est nul ou très restreint. Si ce biais est important, aucune inférence ne peut être faite sur la démographie de la population à l'aide des distributions "mismatch". Il faut préciser que l'étude de SNP dont la fréquence minimum dans la population est de 5% donne des résultats très semblables à ceux obtenus avec une fréquence minimum de 10% (non montré). Nous avons également montré que les SNPs ne permettent pas d'estimer l'âge de l'expansion démographique, puisqu'ils sont indépendants du taux de mutation.

3.3.2 Implications pour les populations européennes

Pereira *et al.* (2001) se sont appuyés sur des distributions "mismatch" tirées de SNPs situés sur le chromosome Y pour émettre l'hypothèse que la lignée masculine européenne ne porte pas la trace d'une expansion démographique. Cette hypothèse a été ensuite étendue au niveau mondial dans un second article, mais sous une forme nuancée puisque Dupanloup *et al.* (2003) supposent qu'une expansion récente s'est produite sans laisser de traces dans les distributions "mismatch". Ces deux études se basent sur les données provenant de 25 populations typées pour 22 SNPs par Semino *et al.* (2000a). Il s'agit de données qui sont sujettes à un biais de recrutement dont l'importance est inconnue. Une inspection visuelle des distributions "mismatch" observées dans les populations européennes et moyen-orientales révèle que la majorité d'entre elles possèdent une homozygotie importante et qu'elles sont toutes multimodales (Figure 3.4). Les simulations que nous avons effectuées nous permettent de réévaluer l'hypothèse émise par Dupanloup *et al.* (2003). En effet, une expansion démographique et spatiale ancienne (4'000 générations) permet d'observer des distributions "mismatch" multimodales du type de celles observées pour le chromosome Y en Europe, soit lorsque le *Nm* de la population est réduit (Figure 3.5a-b), soit lorsqu'il existe un biais de recrutement dans les données (Figure 3.5b-d). Une expansion démographique dans une population de grand *Nm*, sans aucun biais de recrutement, peut, au contraire, être complètement exclue à la vue des distributions "mismatch" produites (Figure 3.5C).

Un second argument en faveur d'une expansion démographique commune à la lignée femelle et à la lignée mâle européenne est l'existence d'une certaine constance dans la forme des distributions "mismatch" observées, qui montrent pour la plupart deux ou trois modes principaux, localisés vers 0, 4 et 8 différences (Figure 3.4). En effet, dans une population stationnaire, on s'attend à observer une variance beaucoup plus grande.

Même si nos arguments ne se fondent que sur une observation visuelle, il ne nous paraît pas possible d'exclure l'hypothèse que la lignée mâle européenne soit passée par une expansion démographique au Paléolithique, comme cela a été proposé par Pereira *et al.* (2001) et Dupanloup *et al.* (2003) à la lueur des distributions "mismatch" simulées ici, et de leur comparaison avec les distributions "mismatch" réelles. A noter que la simulation d'une expansion datée de -100'000 ans à -10'000 ans ne change rien aux distributions "mismatch" obtenues (non montré) puisque les SNPs sont indépendants du taux de mutation. Il nous est donc impossible de dater cette expansion.

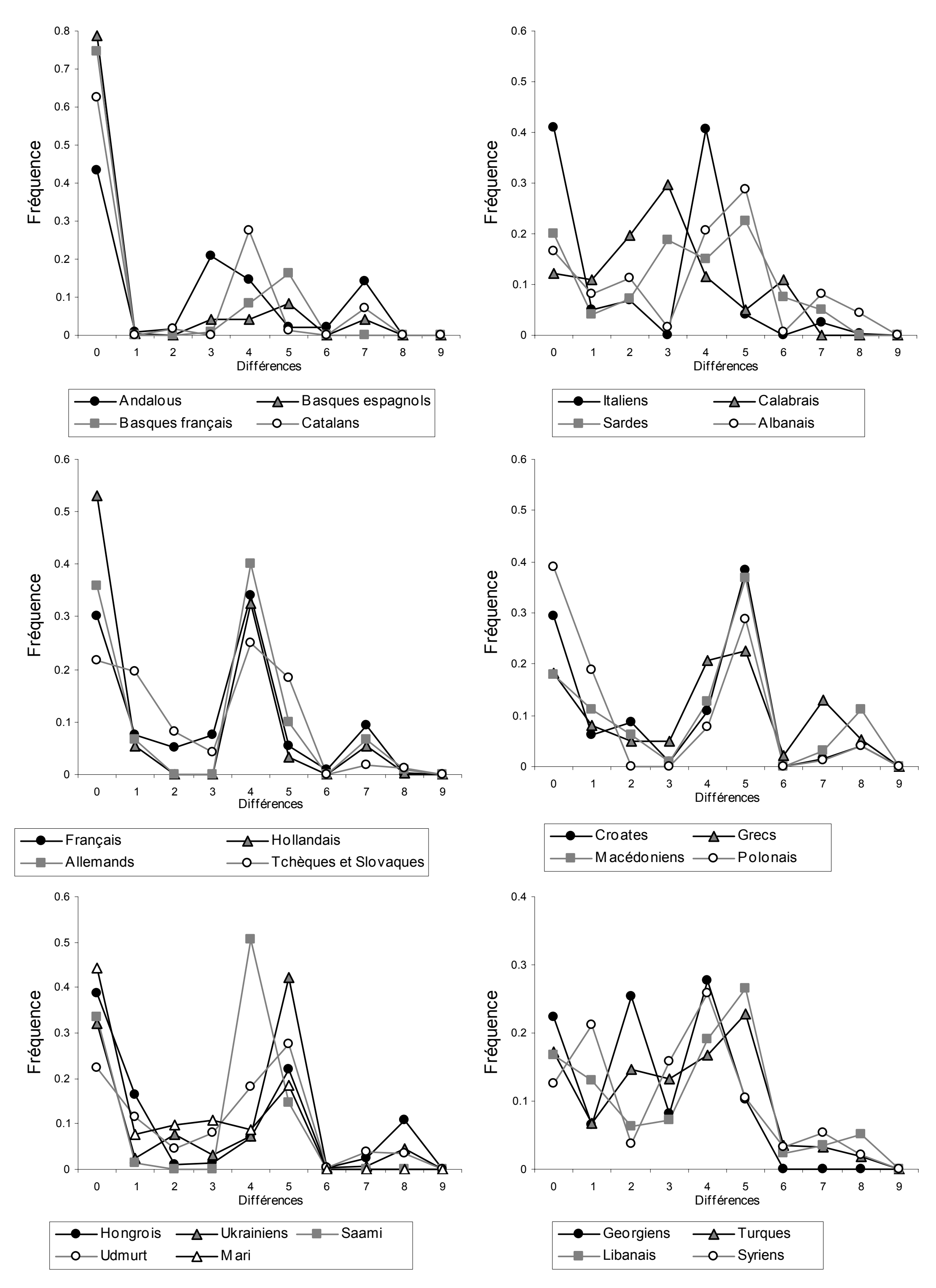


Figure 3.4 Distributions "mismatch" observées dans 25 populations européennes et moyen-orientales, typées pour 22 SNPs (d'après les données de Semino *et al.* 2000a).

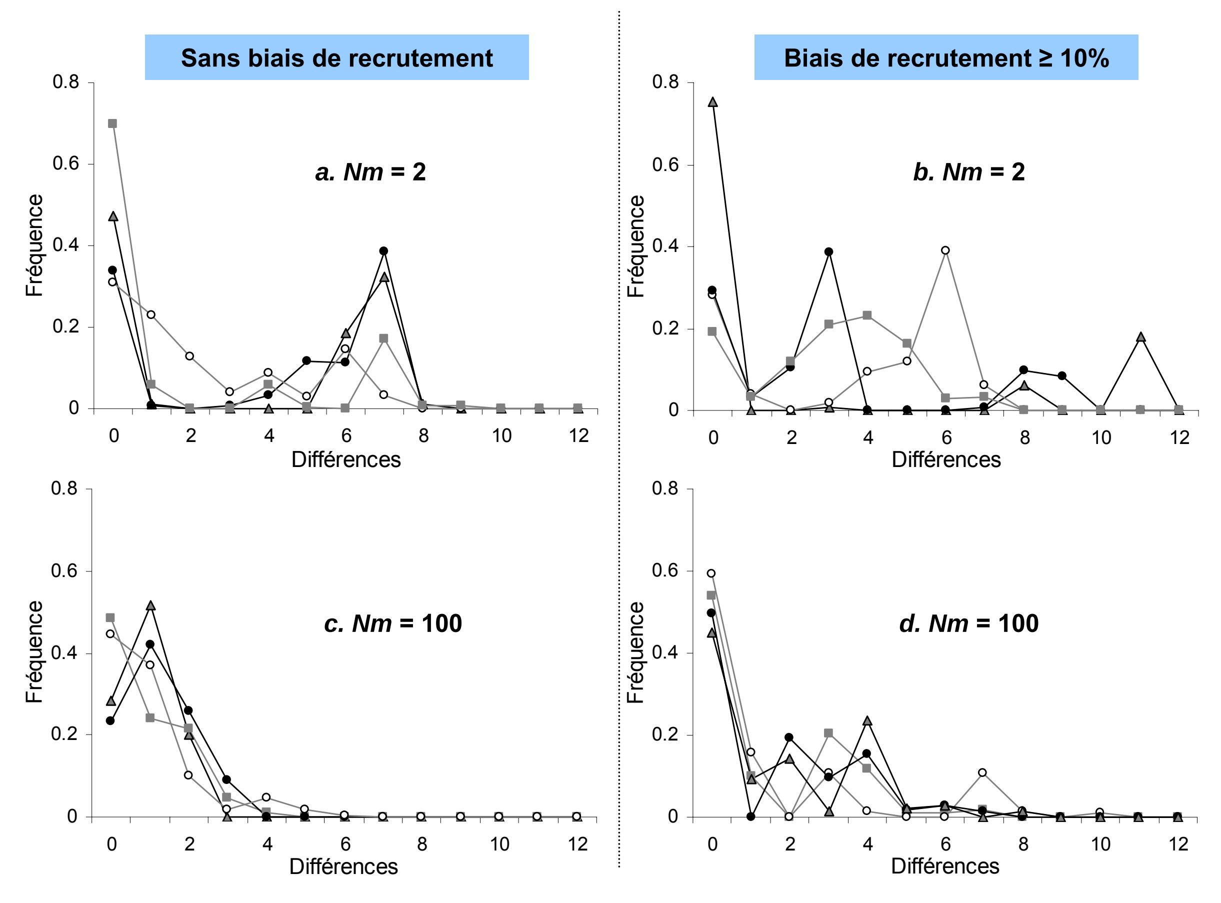


Figure 3.5 Distributions "mismatch" simulées indépendemment avec 11 SNPs dans un échantillon de 40 gènes[1], après une expansion spatiale et démographique dans une population dont le *Nm* est soit petit (ligne du haut), soit grand (ligne du bas). SNPs choisis aléatoirement (colonne de gauche). SNPs avec une fréquence minimum égale à 10% (colonne de droite).

3.4 Discussion

Dupanloup *et al.* (2003) émettent l'hypothèse que les lignées masculines et féminines sont passées par des expansions démographique et spatiale décalées dans le temps. Les femmes auraient connu une expansion plus ancienne que les hommes, dont l'expansion démographique daterait de moins de 10'000 ans lors du passage de la polygamie à la monogamie. Nos simulations ne permettant pas d'écarter l'hypothèse d'une expansion ancienne de la lignée mâle, nous pouvons à notre tour proposer une hypothèse alternative selon laquelle une expansion spatiale et démographique a eu lieu en même temps pour les lignées féminines et masculines, mais que le *Nm* actuel de la population féminine ($N_f m_f$) est plus grand que celui de la population masculine ($N_h m_h$). Un plus grand *Nm* féminin ($N_f m_f > N_h m_h$) peut s'expliquer soit par une plus grande taille efficace *N*, soit par un taux de migration *m* supérieur de la lignée féminine, deux possibilités qui ne sont pas

---

[1] L'effectif moyen des échantillons étudiés par Semino *et al.* (2000) est de 40 individus et le nombre moyen de sites polymorphes par échantillon est égal à 10.7 (+/- 2.5). Comme nous simulons une seule population à la fois, il importe que le nombre de SNPs que nous générons corresponde au nombre de locus effectivement polymorphes dans les échantillons, et non aux 22 SNPs analysés pour l'ensemble des populations.

exclusives. Le premier cas ($N_f > N_h$) est compatible avec une forte polygynie masculine, peu d'hommes ayant beaucoup d'enfants, pouvant résulter d'une polygamie (non-officielle) importante pour les mâles européens actuels. Le second cas, soit un taux de migration des femmes plus grand que celui des hommes ($m_f > m_h$), est une hypothèse qui a déjà été proposée pour les populations humaines (Poloni *et al.* 1997 ; Seielstad *et al.* 1998 ; Oota *et al.* 2002). Ce taux de migration féminin supérieur à celui des hommes peut être expliqué par la patrilocalité[1], qui a vraisemblablement existé dans les populations européennes post-néolithiques (Bentley *et al.* 2002 ; Bentley *et al.* 2003). La transition néolithique aurait donc été le moment à partir duquel la démographie des lignées féminines et masculines se serait différenciée.

Puisque l'importance du biais de recrutement pour le chromosome Y est inconnue, il est donc impossible de comparer directement les distributions "mismatch" observées à l'aide de ce marqueur avec celles observées pour le génome mitochondrial, dont les données sont constituées de séquences complètes. Nous pouvons cependant souligner les ressemblances qui existent entre les distributions "mismatch" mitochondriales des populations Saamis (Figure 3.6) et les distributions tirées du chromosomes Y pour l'ensemble des populations européennes (Figure 3.4). Les Saamis sont des chasseurs-collecteurs, dont l'histoire démographique est différente de celle des populations agropastorales du reste de l'Europe (Sajantila et Paabo 1995 ; Laan et Paabo 1997 ; Kaessmann *et al.* 2002), ce qui peut être traduit par un faible *Nm* dans ces populations (Ray *et al.* 2003). Cette observation est un argument supplémentaire en faveur de l'hypothèse selon laquelle le *Nm* de la lignée masculine européenne est faible, même dans les populations post-néolithiques.

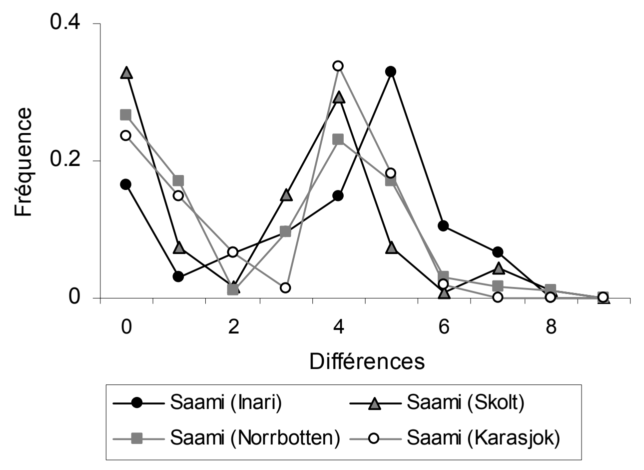


Figure 3.6 Distributions "mismatch" mitochondriales observées indépendamment dans 4 échantillons de Saami provenant de Finlande (Inari et Skolt :Sajantila *et al.* 1995), de Suède (Norrbotten : Sajantila *et al.* 1995) et de Norvège (Karasjok : Delghandi *et al.* 1998).

Nous n'avons malheureusement pas les moyens de tester de façon adéquate cette hypothèse ici, puisque nous ne connaissons pas l'importance du biais de recrutement des données provenant du chromosome Y. Si, dans le futur, il était possible de soustraire ce biais des données du

[1] Une société patrilocale est une société dans laquelle ce sont les femmes qui se déplacent pour se marier, alors que les terres sont héritées par la lignée masculine (voir p. ex. : Oota *et al.* 2001).

chromosome Y, et que la forme des distributions "mismatch" pour ce système restait multimodale, alors il serait possible de soutenir l'hypothèse selon laquelle le *Nm* de la lignée mâle actuel est plus faible que celui de la lignée femelle. Ceci ne peut être réalisé que par le séquençage de portions d'ADN complètes pour la partie non-recombinante du chromosome Y, et pour de nombreuses populations. De telles séquences permettraient non seulement la comparaison entre les signatures moléculaires des lignées féminines et masculines, mais également l'estimation de leur *Nm*. L'utilisation de SPLATCHE, associée à une approche d'estimation bayesienne (p. ex. : Beaumont *et al.* 2001), devrait permettre, dans le futur, de telles estimations pour les populations réelles, humaines ou non.

3.5 Conclusion

Les recherches présentées dans ce chapitre ont permis de montrer que la différence observée dans les distributions "mismatch" des populations de chasseurs-collecteurs contemporains et dans celles des populations post-Néolithique peut être simplement expliquée par une différence de densités. En effet, l'expansion démographique et spatiale d'une population subdivisée dont les dèmes échangent un nombre important de migrants (grand *Nm*) montre une diversité intrapopulationnelle semblable à celle laissée par une croissance démographique instantanée dans une population non-subdivisée. Cette signature est identique à celle observée chez les populations contemporaines qui sont passées par une croissance démographique lors du Néolithique. En revanche, les populations actuelles de chasseurs-collecteurs – dont les densités sont généralement très faibles – montrent une diversité intrapopulationnelle identique à celle laissée par la diffusion d'une population subdivisée dont la combinaison de la taille des dèmes et du taux de migration est faible (petit *Nm*).

Nous avons également montré qu'il n'est pas possible d'exclure une expansion paléolithique de la lignée mâle européenne sur la base des distributions "mismatch" du chromosome Y, contrairement à ce qui a été proposé par Pereira *et al.* (2001) et Dupanloup *et al.* (2003). Cette expansion pourrait avoir été accompagnée d'un taux de migration plus faible pour les hommes que pour les femmes depuis le Néolithique. Cette dernière hypothèse reste cependant à vérifier avec des données sans biais de recrutement portant sur le chromosome Y, principalement à l'aide de séquences d'ADN complètes.

Le logiciel SPLATCHE permet également de comparer les signatures génétiques attendues pour une situation démographique simulée, en fonction de différents types de données (séquence d'ADN, SNP, RFLP, microsatellite, fréquence allélique). Cet aspect peut orienter les recherches futures, en déterminant le type de marqueurs et le nombre de locus nécessaires pour répondre à une question posée. Nous avons ainsi montré que les distributions "mismatch" tirées de SNPs ne sont pas aussi efficaces que celles tirées de séquences d'ADN complètes, pour détecter l'expansion démographique passée d'une population. Au contraire, le biais de recrutement auquel sont sujets les SNPs peut mener à des interprétations erronées des distributions "mismatch" puisqu'il a

tendance à effacer la signature des expansions passées. Il n'est donc pas possible de comparer de façon directe les distributions "mismatch" de la lignée mâle européenne, tirées de SNPs – dont le biais de recrutement est important – de celles de la lignée femelle, tirées de séquences d'ADN complètes. Par conséquent, des séquences d'ADN complètes pour le chromosome Y sont nécessaires à cette comparaison. Ceci souligne l'importance d'avoir des types de données génétiques identiques pour confronter les signatures moléculaires de systèmes différents, entre populations ou entre sexes.

4 Expansion spatiale dans un contexte occupé

4.1 Introduction

Le peuplement préhistorique de l'Europe est marqué par deux transitions démographiques importantes, datées d'environ 45'000 ans et 10'000 ans (Biraben 1979). Outre le début d'une croissance démographique, ces deux événements ont en commun la diffusion de nouvelles technologies depuis le sud-est de l'Europe en direction du nord-ouest du continent (Djindjian *et al.* 1999).

La première transition correspond à l'arrivée des Hommes modernes (*Homo sapiens sapiens*), dans une Europe alors habitée par *Homo neandertalensis*[1]. Ces derniers vont disparaître en moins de 15'000 ans, laissant la place aux Hommes modernes (Mellars 1992 ; Bocquet-Appel et Demars 2000b ; Klein 2003). Si ces deux populations ont coexisté dans certaines régions pendant plusieurs siècles, voire plusieurs millénaires (Stringer et Grun 1991 ; Mellars 1998), l'importance de leurs échanges culturels – et plus particulièrement génétiques – est toujours discutée (Hublin 1988 ; Duarte *et al.* 1999 ; Klein 2003).

La seconde transition correspond à la diffusion des techniques agropastorales depuis le Proche-Orient (Lev-Yadun *et al.* 2000 ; Mazurié de Keroualin 2001). Ces techniques ont été, en partie, amenées en Europe par des nouveaux immigrants issus des premiers agriculteurs proche-orientaux. L'importance de ce flux migratoire est cependant discutée puisque, dans certaines régions, les chasseurs-collecteurs indigènes ont eux-mêmes adopté l'agriculture et ont ainsi participé à la croissance démographique (Harris 1996 ; Whittle 1996 ; Thorpe 1999 ; Mazurié de Keroualin 2003).

Dans les deux cas mentionnés ci-dessus, les données génétiques actuelles devraient permettre d'obtenir des indications quant à la contribution respective des différentes populations au patrimoine génétique européen actuel. Ce sont les questions auxquelles nous nous sommes interessé dans ce travail (chapitres 5 et 6), au moyen de l'approche par simulation présentée dans le chapitre 2. Les deux transitions peuvent ainsi être modélisées par l'expansion spatiale et démographique d'une population (*B*), prenant sa source dans le sud-ouest de l'Asie, dans une aire déjà peuplée par une autre population (*A*). Cette modélisation doit permettre des échanges génétiques, ainsi que de la compétition, entre les deux populations. En effet, lors de leur diffusion en Europe, aussi bien les premiers *Homo sapiens sapiens* que les agriculteurs, ont totalement remplacé les populations qui peuplaient préalablement l'Europe (respectivement les Néandertaliens et les chasseurs-collecteurs). Ces observations laissent raisonnablement penser que les Hommes modernes ont été en compétition avec les Néandertaliens (Bocquet-Appel et Demars 2000b ; Stringer et Davies 2001 ;

---

[1] Il est de plus en plus largement accepté que les Néandertals font partie d'une espèce distincte, appartenant au genre *Homo* (p. ex. Schwartz et Tattersall 1996 ; Tattersall et Schwartz 1999), mais certains auteurs, comme Wolpoff (1996), les considèrent comme une variante de notre propre espèce (*Homo sapiens neandertalensis*). Nous utiliserons la première nomenclature dans ce travail.

Hublin 2002), tout comme les agriculteurs avec les chasseurs-collecteurs (Hyden 1990 ; Spielmann et Eder 1994 : p. 317 ; Van Andel 2000).

Bien que des simulations du remplacement des chasseurs-collecteurs européens par les agriculteurs néolithiques aient déjà été effectuées (voir section 4.2), les programmes utilisés n'ont jamais été mis à la disposition du public. Nous avons donc dû développer notre propre outil de simulation pour étudier cette période de transition, ainsi que celle de l'arrivée des premiers hommes modernes en Europe. Nous avons décidé d'adapter le logiciel SPLATCHE afin que deux populations différentes puissent évoluer simultanément dans une aire géographique virtuelle. La version de base de SPLATCHE présentée dans le chapitre 2 ne permet, en effet, que de simuler la diffusion d'une population unique dans un monde préalablement vide. Parallèlement à la modification de SPLATCHE, il a été également nécessaire de développer notre propre modèle démographique, afin de simuler de manière réaliste les interactions entre deux populations évoluant dans la même aire géographique. En effet, aucun des modèles répertoriés dans la littérature (Rendine *et al.* 1986 ; Barbujani *et al.* 1995 ; Aoki 1996 ; Aoki *et al.* 1996 ; Flores 1998) ne permet de simuler de manière adéquate à la fois des échanges génétiques et de la compétition entre populations.

Dans ce chapitre, nous décrivons donc un modèle démographique qui, implémenté dans SPLATCHE, permet de simuler les interactions entre deux populations en compétition (sections 4.3 et ANNEXE 3). La description est faite de manière très générale, afin que ce modèle, tel quel ou légèrement modifié, puisse être ultérieurement utilisé dans d'autres contextes. Seuls les cas particuliers du modèle utilisés dans les applications présentées dans les chapitres 5 et 6 sont testés ici de façon intensive. Le comportement de notre modèle en fonction de ses différents paramètres est étudié dans un cadre identique à celui utilisé dans le chapitre 3. Cela nous permet de comparer la variabilité génétique obtenue après une expansion démographique et spatiale, soit dans une aire préalablement vide (section 3.2), soit dans une aire déjà peuplée (section 4.5). Nous discuterons les différences observées dans les généalogies et sur les distributions "mismatch" dans les deux situations (4.5.4).

4.2 Différents modèles d'expansion de populations humaines dans une aire occupée

Dans cette section, nous faisons une revue critique de certains modèles d'expansion de populations humaines dans une aire préalablement peuplée qui ont été proposés dans littérature. Il ne s'agit pas d'une revue exhaustive puisque nous présentons, dans l'ordre chronologique de leur publication, uniquement les modèles qui nous paraissent les plus pertinents par rapport à notre propre travail.

- Rendine, Piazza, Cavalli-Sforza : American Naturalist (1986)

Faisant suite à un premier essai dans les années 70 (Sgaramella-Zonta et Cavalli-Sforza 1973), l'étude de Rendine *et al.*, en 1986, aborde la problématique de la transition néolithique en Europe

par le biais de simulations. Cette étude fait suite à une série d'articles (Ammerman et Cavalli-Sforza 1971 ; Menozzi *et al.* 1978 ; Ammerman et Cavalli-Sforza 1984) qui ont permis à ces auteurs de définir la théorie de la "vague d'avancée démique" du Néolithique européen[1]. L'article de Rendine *et al.* (1986) visait à simuler cette vague d'expansion démique afin de confronter la structure génétique obtenue avec celle des populations européennes. L'évolution de fréquences alléliques est simulée dans deux populations, l'une composée de chasseurs-collecteurs mésolithiques (*CC*) et l'autre d'agriculteurs (*AG*). Les simulations se déroulent dans une Europe virtuelle divisées en 840 cellules ou dèmes, placées selon une matrice régulière (Figure 4.1). Les migrations ne peuvent avoir lieu qu'entre dèmes voisins. Il s'agit donc d'un modèle "stepping-stone" en deux dimensions (Kimura 1953). De plus, deux matrices de dèmes sont superposées; elles représentent respectivement la population *AG* ou la population *CC*. Des migrations sont possibles entre dèmes appartenant à des populations différentes si leur localisation est identique dans leur matrice respective. Tous les dèmes de la population *CC* sont peuplés au départ, puis c'est au tour de la population *AG* de coloniser tous les dèmes de la seconde matrice, à partir de six dèmes sources localisés au Proche-Orient. Lors de chaque génération, la démographie des populations *CC* et *AG* se fait selon trois étapes :

1°) Régulation démographique de chacune des deux populations dans chaque dème, selon une croissance logistique définie par l'équation :

$$N_{X,t+1} = N_{X,t} + r_X N_{X,t}(1 - N_{X,t} / K_X) \qquad \text{(Eq. 4.1)}$$

où $N_{X,t}$ est la taille de la population X (avec X = *CC* ou *AG*) à la génération t et r_X son taux de croissance. K_X est la capacité de soutien du dème pour la population X, soit le nombre maximum d'individus qui peuvent vivre simultanément dans un dème en fonction des ressources disponibles.

2°) Migrations intrapopulationnelles vers les 4 dèmes voisins selon un taux m fixe.

3°) Incorporation de S individus *CC* dans la population *AG* (migrations interpopulationnelles, appelées "acculturation") selon l'équation

$$S = \gamma N_{CC,t+1} N_{AG,t+1} \qquad \text{(Eq. 4.2)}$$

Le nombre de migrants interpopulationnels S est simplement fonction du produit des densités de *AG* et de *CC* ainsi que de la constante γ, appelée taux d'acculturation. γ correspond à la probabilité qu'un contact entre un chasseur-collecteur et un agriculteur résulte en l'adoption des techniques

[1] La "vague d'avancée démique" du Néolithique européen est définie comme une vague de migration des premiers agriculteurs depuis le Proche-Orient, à une vitesse constante égale à 1 *Km* par année (Ammerman et Cavalli-Sforza 1984). Selon cette théorie, la contribution des chasseurs-collecteurs indigènes est soit nulle, soit très faible. Cette vague d'avancée ne correspond pas à une migration dirigée, d'un lieu vers un autre, mais à un front de colonisation global créé sous l'effet de nombreuses migrations à courte distance. A chaque étape, la population d'agriculteurs passe par une croissance démographique. Cette théorie a été passablement contestée (Zvelebil 1986 ; Zvelebil et Zvelebil 1988 ; Zvelebil 1989 ; Harpending 2001 ; Mazurié de Keroualin 2001 ; Fiedel et Anthony 2003).

néolithiques par le chasseur-collecteur (acculturation). La valeur de cette constante a été fixée à 0.00024, sans que ne soit jamais mentionnée la source de cette valeur. Selon Cavalli-Sforza lui-même, cette valeur provient d'un congrès à Hawaii dans les années 70 (communication personnelle). Rendine *et al.* ont appliqué leur modèle dans un milieu homogène, où les valeurs de K_{CC} et de K_{AG} sont constantes.

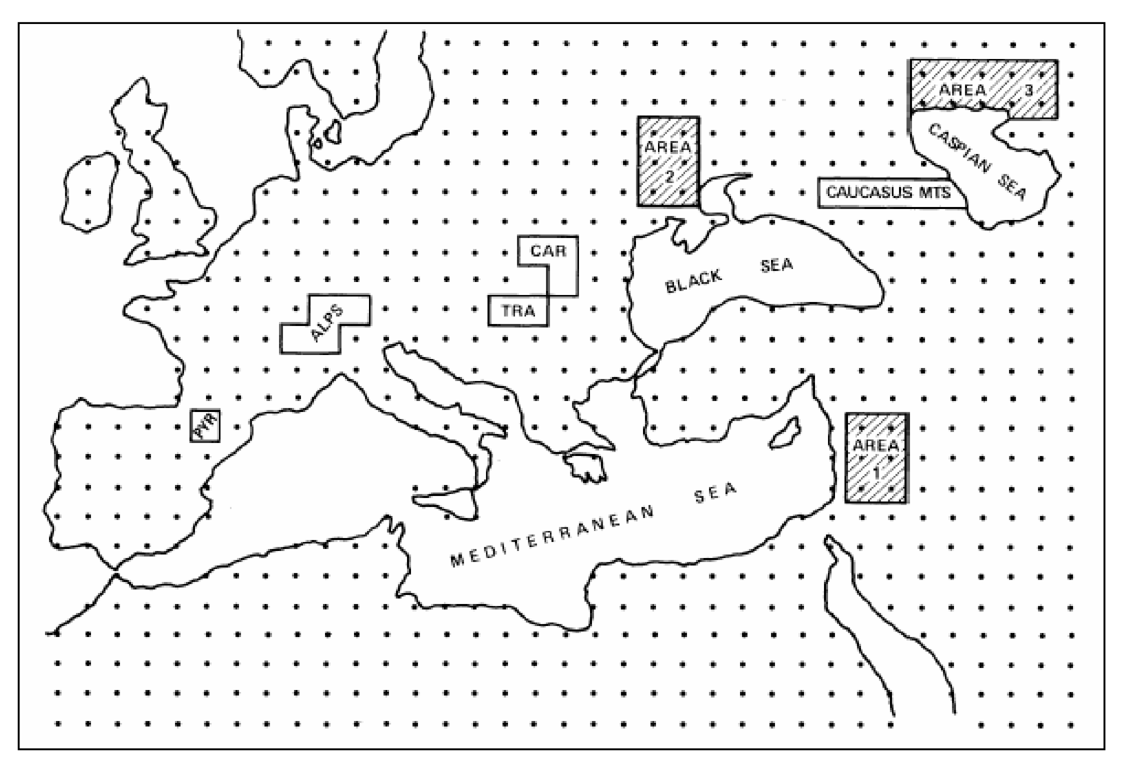


Figure 4.1 Aire virtuelle composée de 840 cellules, dans lesquelles ont lieu les simulations de Rendine *et al.* (1986)

Les inconvénients du modèle proposé par Rendine *et al.* sont les suivants :

i) La condition $N_{AG} \leq 1/\gamma$ doit toujours être respectée pour que le nombre d'émigrants interpopulationnels ne dépasse pas la taille de la population source *CC* (dans le cas contraire, la densité de la population *CC* devient négative !). Les paramètres peuvent êtres aisément choisis pour remplir cette condition dans un monde homogène, lorsque les valeurs de K_{AG} et de γ sont constantes, mais cela devient problématique dans un monde hétérogène ou lorsque différentes combinaisons de paramètres sont utilisées.

ii) Le taux d'acculturation γ (migrations interpopulationnelles) au moment de la transition néolithique est un paramètre inconnu et extrêmement difficile à estimer. Or, dans le travail de Rendine *et al.*, il est fixé à une valeur dont la pertinence est difficilement justifiable, et dont la signification n'est pas claire à nos yeux. Il nous paraîtrait plus adéquat d'utiliser un modèle qui permette de faire varier le nombre de migrations interpopulationnelles afin d'avoir une bonne représentativité de l'influence de ce paramètre.

iii) Le nombre de migrants qui passent par acculturation de la population *CC* vers la population *AG*, par dème et pendant la durée d'une simulation, est d'au moins 350. Il s'agit donc d'une grande

valeur. Si l'on traduit ces chiffres en fonction de la taille des dèmes que nous utilisons dans notre travail (50 *km* de côté au lieu de 156 *km* pour Rendine *et al.*), le nombre de migrants interpopulationnels est d'environ 150 par dème et par simulation. Il serait intéressant de savoir quels seraient les résultats observés en cas d'acculturation plus faible.

iv) La disparition de la population *CC* est effective si la taille des cellules utilisées est identique à celle de Rendine *et al.* (~25'000 km^2). En revanche, avec des cellules dix fois plus petites, comme celles que nous utilisons dans ce travail (2'500 km^2), la population *CC* ne disparaît pas, ce qui ne correspond pas à la réalité.

v) Certaines incohérences existent dans la description du modèle et de certaines valeurs utilisées, par rapport aux équations et aux tables présentées. A titre d'exemple, nous mentionnerons les équations utilisées par Rendine *et al.* pour décrire leur modèle :

$$N_{CC,t+1} = N_{CC,t} + r_{CC} N_{CC,t} \left(1 - \frac{N_{CC,t}}{K_{CC}} - \gamma N_{CC,t} N_{AG,t} \right) \quad \text{(Eq. 4.3)}$$

$$N_{AG,t+1} = N_{AG,t} + r_{AG} N_{AG,t} \left(1 - \frac{N_{AG,t}}{K_{AG}} + \gamma N_{CC,t} N_{AG,t} \right)$$

qui peuvent être reformulées comme :

$$N_{CC,t+1} = N_{CC,t} + r_{CC} N_{CC,t} \left(1 - \frac{N_{CC,t}}{K_{CC}} \right) - \gamma r_{CC} N^2{}_{CC,t} N_{AG,t} \quad \text{(Eq. 4.4)}$$

$$N_{AG,t+1} = N_{AG,t} + r_{AG} N_{AG,t} \left(1 - \frac{N_{AG,t}}{K_{AG}} \right) + \gamma r_{AG} N_{CC,t} N^2{}_{AG,t}$$

Le dernier terme des équations 4.4 correspond à la croissance ou à la décroissance nette d'individus à cause de l'acculturation. On se rend alors compte que le nombre d'individus qui quittent (par acculturation) la population *CC* n'est jamais le même que le nombre d'individus qui arrivent dans la population *AG*, excepté dans le cas particulier où $r_{CC}N_{CC} = r_{AG}N_{AG}$.

Cette incohérence dans la description de cette équation a sans doute échappé à la vigilance des auteurs – il s'agit uniquement d'une parenthèse mal placée – mais elle est loin d'être isolée. Par exemple, les chiffres mentionnés dans le texte et ceux présentés dans les tables sont souvent différents. Toutes ces imprécisions nuisent passablement à la compréhension du modèle. De plus, ce dernier est difficilement applicable dans un cadre général, comme nous l'avons souligné plus haut. Malgré ses points faibles, la méthodologie proposée par Rendine *et al.* (1986) a été très novatrice sous de nombreux aspects, et a eu énormement d'influence sur tous les modèles d'interactions entre chasseurs-collecteurs et agriculteurs qui ont été développés par la suite.

- Calafell, Bertranpetit : Current Anthropology (1993)

La méthodologie de Rendine *et al.* (1986) a été réutilisée par Calafell et Bertranpetit (1993) pour simuler la transition néolithique dans la péninsule ibérique. Cette simulation est cependant plus détaillée, notamment par l'incorporation de plusieurs phases de migrations successives et par

l'utilisation d'une meilleure résolution (réduction de la taille des dèmes). Il est intéressant de noter que le taux d'acculturation est égal à 0.0002 et qu'il a été choisi de façon totalement arbitraire, ce qui souligne une fois de plus l'imprécision autour de ce paramètre.

- Babujani, Sokal, Oden : American Journal of Physical Anthropology (1995)

Barbujani *et al.* (1995) ont également simulé les interactions entre chasseurs-collecteurs paléolithiques et agriculteurs, en développant leur propre cadre de simulation. Le modèle utilisé dans cet article a néanmoins été passablement influencé par celui de Rendine *et al.* (1986) sur de nombreux points. L'évolution de la densité de la population néolithique se fait de la même manière, selon une croissance logistique (équation 4.1), alors que la densité de chasseurs-collecteurs est fixée au départ à 114 individus et diminue uniquement sous l'effet des migrations interpopulationnelles (Figure 4.2). Les migrations interpopulationnelles, apellées acculturation dans cet article, ont la même signification que dans l'étude de Rendine *et al*. (1986) mais sont modélisées de façon différente. En effet, le nombre absolu de migrants *S* qui passent de la population *CC* vers la population *AG* lors de chaque génération est le produit de la probabilité de contact entre chasseurs-collecteurs et agriculteurs et de la probabilité qu'une acculturation découle d'un tel contact:

$$S = N_{AG}\gamma \frac{2N_{AG}N_{CC}}{\left(N_{AG} + N_{CC}\right)^2} \qquad \text{(Eq. 4.5)}$$

Barbujani *et al.* (1995) ont introduit un surplus de réalisme en utilisant des probabilités de migrations différentes, dues aux barrières géographiques telles que les montagnes ou les mers. Cependant, seuls 3 types d'environnements sont pris en compte (Plaine, Montagne et Mer). Ce concept est néanmoins très intéressant puisqu'un individu aura plus ou moins de chances de migrer dans un dème voisin en fonction du type d'environnement de celui-ci.

Les inconvénients du modèle de Barbujani *et al.* (1995) sont les suivants :

i) La population *CC* ne disparaît jamais puisqu'en 440 générations simulées l'amplitude de *S* n'est pas suffisante pour mener *CC* à une extinction. Or, la disparition des chasseurs-collecteurs en Europe est un fait avéré.

ii) La valeur de la probabilité d'acculturation γ est reprise de l'article de Rendine *et al.* (1986) et est gardée constante pour toutes les simulations, sa valeur est égale à 0.00024. Le monde étant homogène (K_{CC} vaut 114 et K_{AG} vaut 7560), *S* est toujours inférieur à 0.1. Le taux d'acculturation est donc très faible, le maximum étant d'un migrant toutes les 10 générations. Comme la population *CC* ne disparaît jamais, cette faible probabilité d'acculturation permet tout de même d'observer au maximum une quarantaine de migrations par dème pendant la totalité des 440 générations simulées. Selon Barbujani, le nombre de migrants par acculturation devrait plutôt approcher 1 individu par génération à la lumière des données récentes (communication personnelle).

iii) Le problème de la signification exacte de γ et de sa valeur subsiste. De même que l'impossibilité de faire varier avec une amplitude satisfaisante le nombre de migrants interpopulationnels *S*.

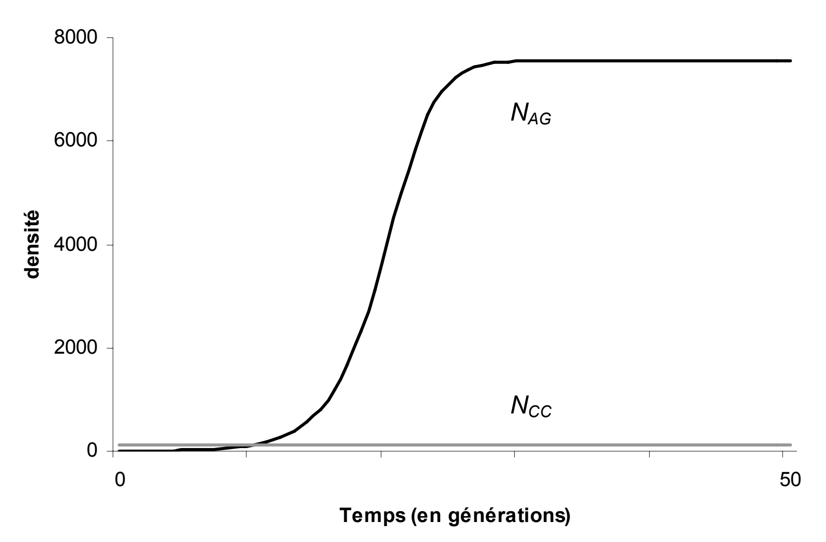


Figure 4.2 Evolution des densités de chasseurs-collecteurs (en gris) et d'agriculteurs (en noir), au cours du temps avec les valeurs de paramètres tirés de Barbujani *et al* (1995). Soit K_{CC} = 114, K_{AG} =7560, r_{AG} = 0.5 et γ = 0.00024. Trait fin = *S* mais est toujours très petit et donc indiscernable sur cette figure.

Bien que le modèle de Barbujani *et al.* (1995) soit le plus détaillé réalisé à ce jour, certains aspects paraissent assez peu réalistes, notamment la dynamique des chasseurs-collecteurs, qui est tout simplement absente.

- Aoki : World Scientific (1996)

Aoki (1996) a également abordé la modélisation des interactions entre chasseur-cueilleurs et néolithiques, mais de manière analytique. Son modèle utilise l'apport d'une troisième population *AC* qui correspond aux chasseurs-collecteurs convertis aux techniques néolithiques. Aoki fait donc la distinction entre la population néolithique issue de la population d'origine (*AG*) et celle (*AC*) constituée des néolithiques descendant, par assimilation, des chasseurs-collecteurs. Le modèle utilisé peut-être décrit comme :

$$N_{CC,t+1} = N_{CC,t} + \alpha_{CC} N_{CC,t}\left(1 - \frac{N_{CC,t}}{K_{CC}}\right) - \gamma N_{CC,t}(N_{AG,t} + N_{AC,t}) \qquad \text{(Eq. 4.6)}$$

$$N_{AC,t+1} = N_{AC,t} + \alpha_{AC} N_{AC,t}\left(1 - \frac{\left(N_{AG,t} + N_{AC,t}\right)}{K_{AG}}\right) + \gamma N_{CC,t}(N_{AG,t} + N_{AC,t})$$

$$N_{AG,t+1} = N_{AG,t} + \alpha_{AG} N_{AG,t}\left(1 - \frac{\left(N_{AG,t} + N_{AC,t}\right)}{K_{AG}}\right)$$

où N_{AC} est la densité de chasseurs-collecteurs convertis aux techniques néolithiques.

Le modèle développé par Aoki est difficilement comparable au nôtre pour plusieurs raisons :

i) Son modèle ne considère pas de mariages mixtes entre chasseurs-collecteurs et agriculteurs. En effet, les néolithiques convertis restent dans une population séparée des néolithiques d'origine. Si cette subdivision de la population des agriculteurs peut s'expliquer par une simplification du modèle mathématique, en revanche, d'un point de vue génétique la dynamique d'une population subdivisée est différente de celle d'une population non subdivisée. De plus, l'utilisation de la théorie de la coalescence (comme dans notre approche, voir Annexe 2.2) ne nous oblige pas à invoquer l'existence d'une telle population *AC*.

ii) L'influence de l'environnement n'est pas prise en compte.

iii) Le modèle de Aoki simule de la compétition uniquement entre Néolithiques convertis (*AC*) et Néolithiques originaux (*AG*), mais pas entre agriculteurs et chasseurs-collecteurs (*CC*), qui sont considérés comme évoluant dans des aires différentes. Il n'existe aucun type d'interactions entre agriculteurs originaux et chasseurs-collecteurs.

4.3 Modèle démographique proposé

Nous avons développé notre propre modèle démographique pour simuler l'expansion d'une population dans une aire déjà peuplée. Ce modèle s'inspire de ceux qui ont été présentés dans la section précédente (4.2) mais offre de nombreux avantages que nous mentionnerons dans la section 4.4. Il permet de simuler non seulement la diffusion de deux populations dans l'espace, mais également des échanges génétiques et de la compétition entre elles.

La dynamique des deux populations, que nous nommerons *A* et *B*, peut être décomposée en deux parties : *i)* régulation démographique intra-dème, et *ii)* migrations d'individus entre dèmes voisins (de la même population ou entre populations différentes).

4.3.1 Régulation démographique intra-dème

Sans tenir compte d'éventuelles migrations et sous l'hypothèse de conditions environnementales stables, les variations de la densité d'une population au cours du temps dépendent de la différence entre le taux de natalité et le taux de mortalité. Cette différence peut être résumée par *r*, le taux intrinsèque de croissance (Begon *et al.* 1996 : p. 165). En théorie, une population qui vit dans un milieu sans aucune contrainte suivrait un accroissement démographique exponentiel infini du type :

$$N_{t+1} = N_t + r N_t \qquad \text{(Eq. 4.7)}$$

où N_t est égal à la densité au temps *t* (Begon *et al.* 1996: p. 246).
Lorsque $r > 1$, la population est en croissance, et lorsque $r < 1$, elle décroît.

Or, ce modèle ne correspond pas aux populations réelles puisque la croissance démographique de celles-ci est limitée par la quantité de ressources disponibles, qui ne sont évidemment pas infinies. Une compétition pour ces ressources (qu'il s'agisse de nourriture, d'espace, d'habitat, ou

autres) se met donc en place au cours du temps. Cette compétition peut s'exercer entre individus de la même population (compétition intrapopulationnelle) ou entre individus de populations différentes, qui occupent la même zone (compétition interpopulationnelle). Sous l'effet de la compétition, les taux de natalité et de mortalité varient au cours du temps, au fur et à mesure des variations de densité.

4.3.1.1 Compétition intrapopulationnelle

Le modèle de croissance exponentielle (équation 4.7) ne tient pas compte de la réduction du taux de croissance dû à la compétition intraspécifique pour les ressources environnementales. Verhulst, en 1838, a introduit l'équation de croissance logistique[1], défini par

$$N_{t+1} = N_t\left(1 + r\left(1 - \frac{N_t}{K}\right)\right) \qquad \text{(Eq. 4.8)}$$

où K (la capacité de soutien) représente le nombre maximum d'individus qui peuvent vivre à l'équilibre dans une aire donnée (Begon *et al.* 1996: p. 247).

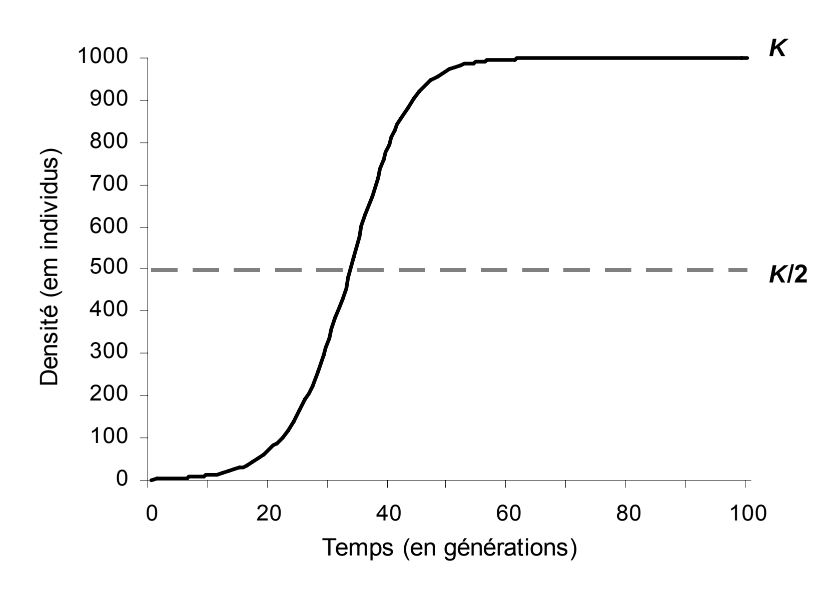


Figure 4.3 Exemple d'une croissance logistique avec K = 1'000 et r = 0.2.

Cette équation est une extension du modèle de croissance exponentielle, mais elle tient compte du niveau de saturation caractéristique de l'environnement K (Begon *et al.* 1996: p. 224). Aux densités inférieures à K, le taux de natalité excède le taux de mortalité et la taille de la population augmente. Aux densités supérieures à K, le taux de mortalité excède le taux de natalité et la population diminue. K représente donc un équilibre. Le dernier terme de l'équation 3.2 permet de réguler l'accroissement intrinsèque de la population. En effet, plus la densité de la population N se rapproche de la densité maximum du milieu K et plus la croissance sera faible, pour être finalement nulle lorsque la capacité du milieu est atteinte.

---

[1] Voir Tsoularis et Wallace (2002) pour une revue des différents types d'équations logistiques.

Les trois caractéristiques principales de l'équation logistique sont (Figure 4.3) :

i) $\lim_{t \to \infty} N_t = K$, la population tend vers sa capacité de soutien.

ii) Le taux de croissance relatif décroît linéairement avec l'accroissement de la densité et atteint 0 lorsque $N = K$.

iii) Au point d'inflexion, N est égal à la moitié de K ($N=K/2$) et le taux de croissance absolu est à son maximum et vaut $rK/4$.

4.3.1.2 Compétition interpopulationnelle

Afin de modéliser la compétition entre deux populations, nous avons décidé d'utiliser le modèle de Lotka-Volterra, qui est le modèle classique de compétition interspécifique en écologie. Il offre les avantages d'être relativement facile à comprendre et à implémenter. De plus, il permet de modéliser de manière satisfaisante les comportements désirés, comme nous le verrons ci-dessous.

Lotka et Volterra ont défini un modèle de compétition interspécifique, qui est une extension du modèle de croissance logistique (Volterra 1926 ; Lotka 1932). Ce modèle inclut à la fois les effets de la compétition intrapopulationnelle et ceux de la compétition entre les populations A et B:

$$N_{A,t+1} = N_{A,t}\left(1 + r_A \frac{\left(K_A - N_{A,t} - \alpha_{AB} N_{B,t}\right)}{K_A}\right) \qquad \text{(Eq. 4.9)}$$

$$N_{B,t+1} = N_{B,t}\left(1 + r_B \frac{\left(K_B - N_{B,t} - \alpha_{BA} N_{A,t}\right)}{K_B}\right)$$

où le coefficient de compétition α_{AB} représente l'effet que la population B exerce sur la population A (Begon *et al.* 1996: pp. 274-278).

Le terme $-\alpha_{AB}N_{B,t}$ diminue le taux d'accroissement de la population A au fur et à mesure que la densité de la population B, en compétition, augmente. Cette diminution est plus ou moins forte, selon l'importance de la compétition interpopulationnelle, représentée par α. On peut voir α_{AB} comme l'effet inhibiteur d'un individu de la population B sur un individu de la population A, comparé à l'effet inhibiteur d'un individu de la population A sur un autre individu de la même population A.
α peut prendre différentes valeurs :

- si $\alpha_{AB} = 0$, cela signifie que la population B n'exerce aucune compétition sur la population A, donc pas de compétition interpopulationnelle dans ce sens.
- si $\alpha_{AB} = 1$, cela signifie qu'un individu de la population B exerce autant de compétition sur un individu de la population A qu'un autre individu de cette même population A. En d'autres termes, que la compétition interpopulationnelle est égale à la compétition intrapopulationnelle, ou que la compétition interpopulationnelle est complète.
- si $\alpha_{AB} < 1$, cela signifie que la compétition intrapopulationnelle est plus forte que la compétition interpopulationnelle et donc qu'un individu de la population B a un effet inhibiteur sur un individu de la population A, qui est plus faible que celui exercé par un autre individu de la population A.

- si $\alpha_{AB} > 1$, cela signifie que la compétition interpopulationnelle est plus forte que la compétition intrapopulationnelle et donc qu'un individu de la population *B* a un effet inhibiteur sur un individu de la population *A,* qui est plus fort que celui exercé par un autre individu de la population *A*. Ce cas de figure est assez rare dans la réalité.

La compétition entre deux populations peut-être symétrique, mais elle est le plus souvent asymétrique, les taux de compétition dans un sens ou dans l'autre étant différents ($\alpha_{AB} \neq \alpha_{BA}$).

Quatre cas différents peuvent se présenter selon les valeurs que prennent les variables K_A, K_B, α_{AB} et α_{BA} de ce modèle.

1. Si $K_A > K_B\alpha_{AB}$ et $K_B < K_A\alpha_{BA}$, la population *B* va finalement disparaître à cause de la compétition exercée sur elle par la population *A*.
2. Si $K_A < K_B\alpha_{AB}$ et $K_B > K_A\alpha_{BA}$, cas inverse au précédent, la population *A* va finalement disparaître à cause de la compétition exercée sur elle par la population *B*.
3. Si $K_A < K_B\alpha_{AB}$ et $K_B < K_A\alpha_{AB}$, les deux populations subissent une compétition interpopulationnelle plus forte que la compétition intrapopulationnelle. La population avec l'effectif initial le plus grand va pousser l'autre population à l'extinction.
4. Si $K_A > K_B\alpha_{AB}$ et $K_B > K_A\alpha_{BA}$, les deux espèces subissent une compétition intrapopulationnelle plus forte que la compétition interpopulationnelle qui s'exerce entre elles. Un équilibre stable s'établit alors.

Nous reviendrons plus tard sur ces 4 cas généraux, en fonction des différents modèles spécifiques que nous allons décrire.

Il faut relever que l'état final du système ne dépend que des variables *K* et α, mais pas de *r*, puisque ce facteur influence seulement la rapidité avec laquelle l'état final se met en place.

4.3.1.3 Modèles de compétition développés

Nous avons jusqu'à présent fait une revue du modèle de Lotka-Volterra, tel qu'il est décrit dans la littérature. Sur cette base, nous avons développé quatre modèles différents de compétition qui peuvent s'appliquer spécifiquement au remplacement des Néandertaliens et à la transition néolithique en Europe. Nous allons les décrire, puis les comparer.

- Modèle M1 : Taux de compétition fixés

Dans la situation spécifique pour laquelle une des deux populations disparaît au cours du temps, nous pouvons raisonnablement penser que la compétition interpopulationnelle est asymétrique. L'hypothèse de base est que, dans les environnements favorables à la population *B*, son avantage compétitif est complet sur la population *A* ($\alpha_{AB} = 1$), tandis que dans ces mêmes milieux les individus *A* n'exercent pas d'influence sur les individus *B* ($\alpha_{BA} = 0$). Cette hypothèse est assez grossière mais elle paraît raisonnable et permet de modéliser de façon satisfaisante le

remplacement d'*Homo Neandertalensis* par les premiers *Homo sapiens sapiens*, puis le remplacement des chasseurs-collecteurs paléolithiques par les agriculteurs néolithiques. Le modèle "taux de compétition fixé" peut s'écrire comme :

$$N_{A,t+1} = N_{A,t}\left(1 + r_A \frac{\left(K_A - N_{A,t} - N_{B,t}\right)}{K_A}\right) \qquad \text{(Eq. 4.10)}$$

$$N_{B,t+1} = N_{B,t}\left(1 + r_B \frac{\left(K_B - N_{B,t}\right)}{K_B}\right)$$

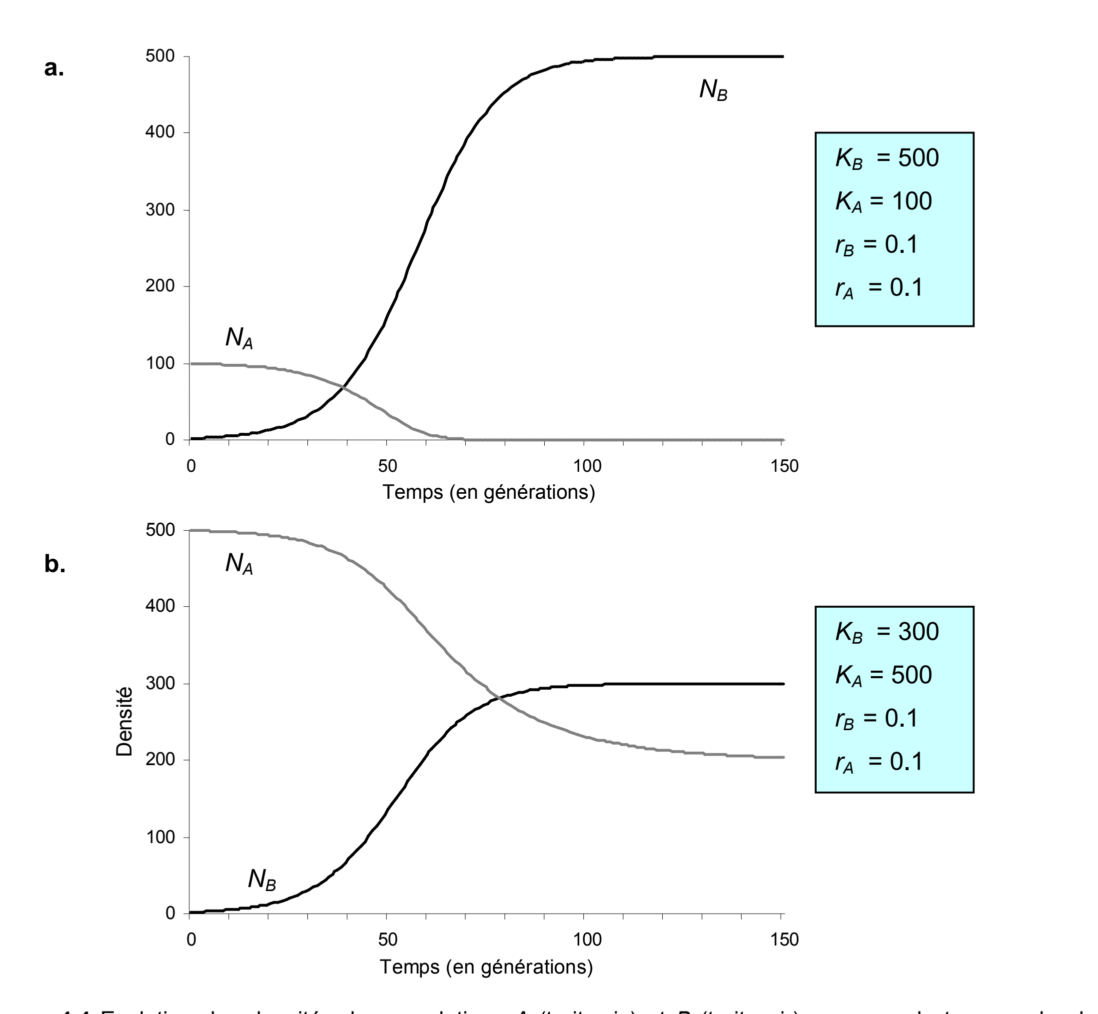


Figure 4.4 Evolution des densités des populations *A* (trait gris) et *B* (trait noir) au cours du temps selon le modèle 1 (Lotka-Volterra avec taux de compétition fixés comme α_{BA} = 0 et α_{AB} = 1). a : Disparition de la population *A* car $K_B > K_A$. b : Equilibre stable entre les deux populations car $K_B < K_A$.

Dans cette application précise, les cas généraux 2 et 3 du modèle de Lotka-Volterra (voir page 51) sont impossibles car K_B n'est jamais négatif. Dans tous les cas, la population *B* qui colonise une cellule, s'y établit de façon durable (cas 1 ou 4). La population *A* disparaît seulement si $K_B > K_A$ (cas 1, Figure 4.4a). Dans le cas contraire, un équilibre stable entre les deux populations s'établit (Figure

4.4b). Dans tous les cas, la capacité de soutien K de la cellule est égale à la capacité de soutien la plus élevée des deux populations ($K = \max(K_A, K_B)$).

Ce modèle, utilisant des taux de compétition fixés *a priori*, n'est pas applicable de façon générale dans un environnement hétérogène. Par exemple, dans certains milieux spécifiques la population A peut être favorisée, alors que dans d'autres milieux c'est la population B qui est favorisée. Dans un environnement hétérogène les relations de compétition sont différentes d'un milieu à l'autre et ne peuvent donc pas être fixées une fois pour toutes.

- Modèle M2 : Taux de compétition dépendant de la densité

Afin d'éviter le choix *a priori* des taux de compétition entre populations, nous avons développé un autre modèle, plus général, pour lequel α_{AB} et α_{BA} ne sont pas fixés au départ, mais varient au cours du temps en fonction des densités de population. Ceci reflète le fait que la population dont la densité est la plus forte dans une cellule, au moment t, est celle qui exerce le plus de compétition sur l'autre population. Ainsi, nous avons considéré que :

$$\alpha_{AB,t} = \frac{N_{B,t}}{(N_{B,t} + N_{A,t})} \qquad \text{(Eq. 4.11)}$$

$$\alpha_{BA,t} = \frac{N_{A,t}}{(N_{B,t} + N_{A,t})}$$

avec $\alpha_{AB,t} + \alpha_{BA,t} = 1$.

Dans ce modèle, l'état final du système dépend uniquement des valeurs de K_A et K_B. Etudions les différents cas possibles lors de l'arrivée d'individus de la population B dans une cellule déjà peuplée par la population A :

1. $K_A < K_B$: Il s'agit du cas général n°2 du modèle de Lotka-Volterra (page n° 51). Il en résulte un envahissement de B et une extinction de A (Figure 4.5a).
2. $K_A = K_B$: Il s'agit du cas général n°4 du modèle de Lotka-Volterra. Il en résulte un équilibre stable, A et B cohabitent indéfiniment (Figure 4.5b).
3. $K_A > K_B$: Il s'agit du cas inverse au premier ($K_B > K_A$), soit le cas général n°1 du modèle de Lotka-Volterra. Il en résulte que B ne parvient pas à envahir et disparaît rapidement.

Le cas général n°3 du modèle de Lotka-Volterra n'est jamais possible, puisque $\alpha_{AB,t} + \alpha_{BA,t} = 1$, la compétition interpopulationnelle est donc toujours plus faible que la compétition intrapopulationnelle.

Si l'état final du système ne dépend que de K_A et K_B, en revanche, la vitesse à laquelle il se met en place dépend à la fois des taux de croissance r_A et r_B, ainsi que des densités initiales des deux populations ($N_{A,0}$ et $N_{B,0}$, soit les densités au moment où les premiers individus de la seconde population (B) arrivent dans la cellule).

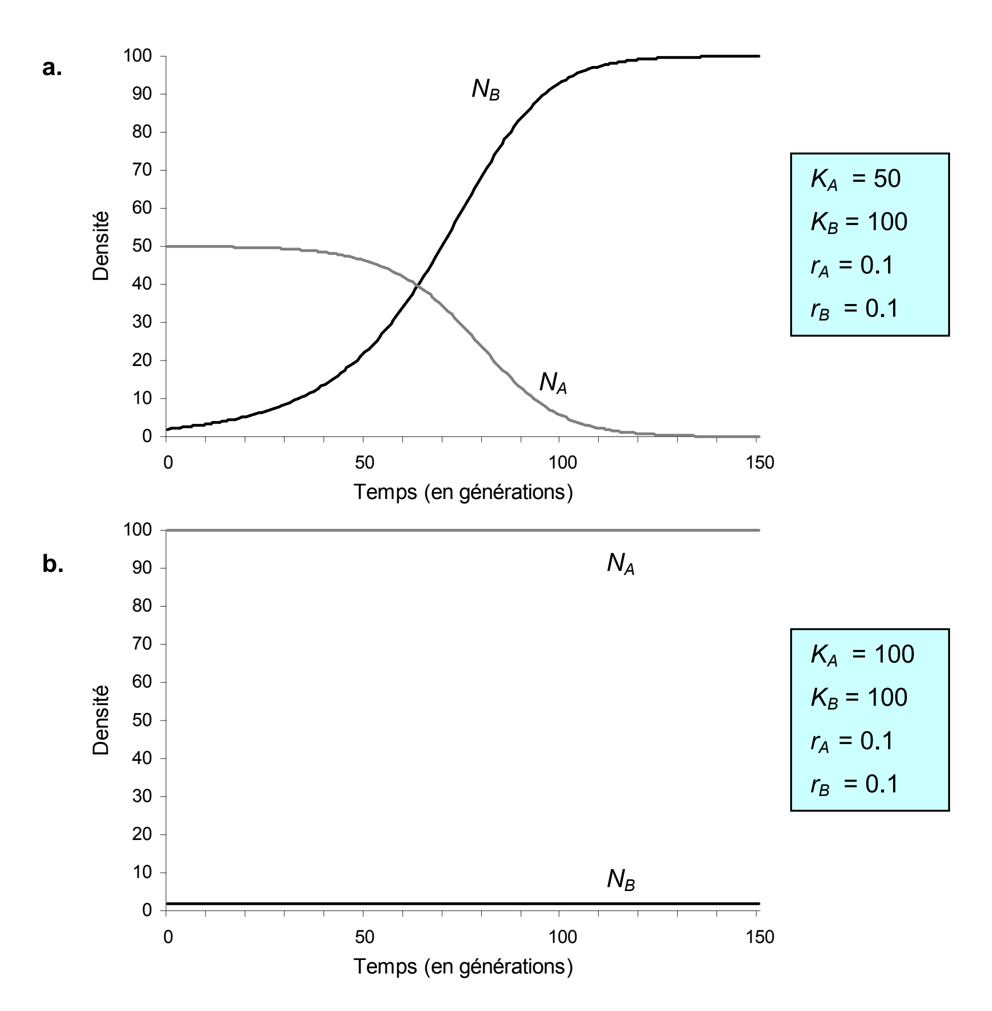


Figure 4.5 Evolution des densités des populations *A* (trait gris) et *B* (trait noir) au cours du temps selon le modèle 2 (Lotka-Volterra avec taux de compétition dépendants des densités). a : Disparition de la population *A* car $K_B > K_A$. b : Equilibre stable entre les deux populations car $K_B = K_A$. La densité initiale de *B* ($N_{B,0}$) est égale à 2 individus.

- Modèle M3 : Cohabitation forcée puis compétition avec taux fixés

Il est possible que la compétition entre deux populations ne se déroule pas directement lors de l'arrivée de la seconde dans une aire déjà peuplée. Par exemple, certains auteurs pensent que les néolithiques ne sont pas directement entrés en compétition avec les chasseurs-collecteurs lorsqu'ils se sont installés dans les mêmes régions (Pinhasi *et al.* 2000). L'installation primaire des communautés néolithiques se serait faite dans des zones différentes de celles occupées par les chasseurs-collecteurs, dans les zones les plus favorables à l'agriculture et à l'élevage. Ce ne serait que dans une seconde phase, lorsque leurs effectifs auraient augmenté, que les agriculteurs auraient commencé à empiéter sur les territoires des chasseurs-collecteurs et que la compétition interpopulationnelle aurait alors débuté.

Dans l'optique de modéliser cette période de cohabitation pendant laquelle aucune compétition ne s'exerce entre les deux communautés, nous avons développé le modèle suivant :

$$N_{A,t+1} = \begin{cases} N_{A,t}\left(1 + r_A \dfrac{(K_A - N_{A,t})}{K_A}\right) & si\ N_{B,t} < K_B - K_A \\ N_{A,t}\left(1 + r_A \dfrac{(K_A - N_{A,t} - N_{B,t})}{K_A}\right) & si\ N_{B,t} \geq K_B - K_A \end{cases} \quad \text{(Eq. 4.12)}$$

$$N_{B,t+1} = N_{B,t}\left(1 + r_B \frac{(K_B - N_{B,t})}{K_B}\right)$$

Selon ce modèle, K_B est égale à la capacité de soutien totale de la cellule. Tant que N_B n'a pas atteint la valeur K_B - K_A, aucune compétition ne s'exerce et la dynamique des deux populations s'établit selon une croissance logistique simple. En revanche, dès que N_B atteint la valeur $K_B - K_A$, alors la compétition interpopulationnelle entre en jeu et la dynamique des populations se fait selon le modèle de Lotka-Volterra (Figure 4.6).

Lorsque $K_B \leq K_A$, ce modèle de cohabitation est identique au modèle de Lotka-Volterra puisque la condition $N_B \geq K_B$ - K_A est toujours respectée. Dans ce cas, l'état final est un équilibre entre les deux populations. En revanche, lorsque $K_B > K_A$, la population *A* disparaît (Figure 4.6).

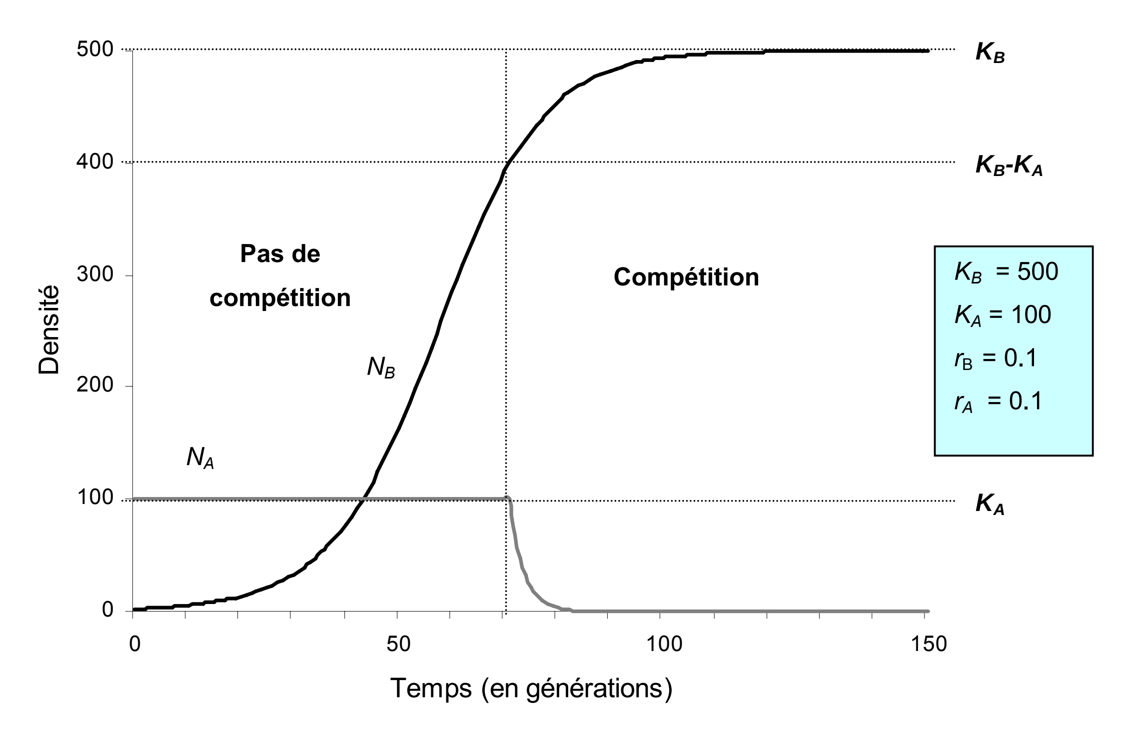


Figure 4.6 Evolution des densités des populations *A* (trait gris) et *B* (trait noir), au cours du temps, selon le modèle M3 (Lotka-Volterra, avec taux de compétition fixés comme $\alpha_{BA} = 0$ et $\alpha_{AB} = 1$ et temps de cohabitation forcée). La population *A* disparaît car $K_B > K_A$.

- Modèle M4 : Cohabitation forcée puis compétition avec taux dépendants de la densité

Le modèle de cohabitation forcée *M3* peut également être implémenté avec des taux de compétition qui dépendent des densités. Tout comme pour le modèle *M3*, aucune compétition ne s'exerce lorsque N_B est inférieur au seuil K_B - K_A. Une fois ce seuil atteint, les deux populations entrent en compétition. A la différence du modèle *M3*, les taux de compétition ne sont pas fixés mais dépendent des densités respectives des deux populations et sont définis par l'équation 4.11, tout comme pour le modèle *M2*.

Avec ce modèle, lorsque $K_B < K_A$, ce modèle de cohabitation est identique au modèle de Lotka-Volterra puisque la condition $N_B \geq K_B - K_A$ est toujours respectée et par conséquent la population *B* ne peut pas s'établir dans la cellule et disparaît. Un équilibre stable s'installe lorsque $K_B = K_A$, alors que la population *A* disparaît si $K_B > K_A$ (Figure 4.6).

4.3.1.4 Comparaison entre les modèles de compétition proposés

Dans la section précédente (4.3.1.3), nous avons défini 4 modèles démographiques (*M1-M4*) qui décrivent l'évolution au cours du temps de deux populations en compétition pour les mêmes ressources. Il importe maintenant de comparer ces 4 modèles entre eux, afin de sélectionner le (ou les) plus adapté(s) à nos recherches, et ainsi de réduire le nombre de simulations à effectuer ultérieurement.

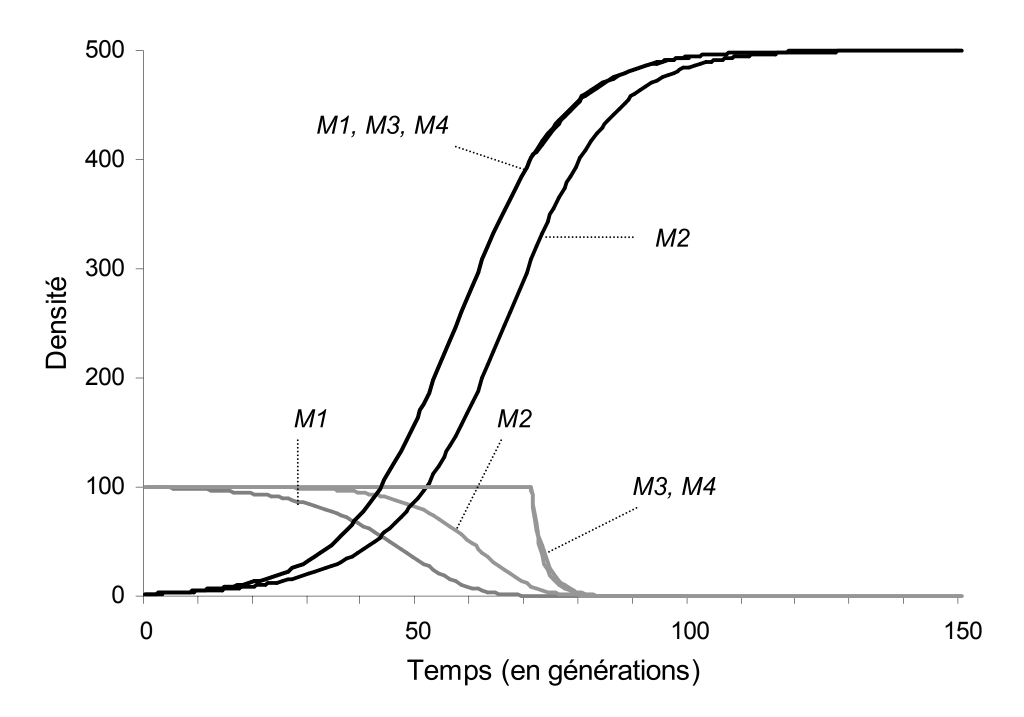


Figure 4.7 Evolution de la densité des populations *A* (en gris) et *B* (en noir), pendant 150 générations, selon 4 modèles démographiques différents (*M1*, *M2*, *M3*, *M4*). Apparition de 2 individus *B* au temps 0, alors que 100 individus *A* peuplent déjà la cellule. K_A = 100, K_B = 500, $r_C = r_B$ = 0.1.

La comparaison entre ces différents modèles (Figure 4.7) lorsque $K_B > K_A$ – qui est le cas particulier qui nous intéresse dans ce travail – montre que :

1. Les modèles *M3* et *M4* sont quasiment identiques, puisque la compétition dépendante de la densité (*M4*) commence à s'exercer à un moment où la densité N_B est très importante, et donc où α_{AB} se rapproche de 1, qui est la valeur fixée pour le modèle *M3*.
2. Le temps de cohabitation des deux populations dans une cellule est plus faible pour les modèles qui ne comprennent pas de période de cohabitation forcée (*M1* et *M2*). Cependant la réduction de ce temps de cohabitation est nettement moins marquée pour le modèle *M2* que pour le modèle *M1*.
3. Le modèle *M2* est intermédiaire entre le modèle *M1* et les modèles *M3* et *M4*.
4. Seul le modèle *M2* influence le temps de colonisation de la population *B* en le ralentissant. En effet, sous ce modèle, la population qui arrive en second (*B*) dans la cellule reste plus longtemps à de faibles densités. Ce qui ralentit sa dispersion dans les autres cellules.

On peut dire que les 4 modèles présentés ici montrent globalement les mêmes caractéristiques, avec des temps de cohabitation qui ne varient que faiblement d'un modèle à l'autre. De tels écarts peuvent cependant être obtenus indépendamment avec n'importe lequel des quatre modèles, en faisant varier leurs paramètres démographiques. Les modèles qui utilisent des taux de compétition dépendant des densités (*M2* et *M4*) possèdent un avantage sur les modèles qui utilisent des taux de compétition fixés (*M1* et *M3*), puisque ces derniers impliquent un jugement de valeur *a priori* sur les populations. Nous avons donc décidé de ne pas retenir les modèles *M1* et *M3* pour la suite de cette étude.

L'ajout d'un temps de cohabitation forcé entre les deux populations peut se justifier d'un point de vue historique. Cependant, les temps de cohabitation effectifs observés pour le modèle *M2* se rapprochent passablement de ceux observés pour le modèle *M4*, mais le *M2* possède l'avantage de simuler une disparition moins abrupte de la population *A*, ce qui semble plus réaliste. De plus, lorsque $K_B > K_A$ le modèle *M4* se comporte quasiment comme le *M3*. *M4* n'utilise donc pratiquement pas la capacité de variation de taux de compétition dépendant de la densité. Par conséquent, nous avons décidé d'écarter également ce modèle au profit du *M2*, afin de simplifier le nombre de cas à étudier lors de nos analyses ultérieures.

Le modèle de compétition *M2* retenu présente l'avantage de pouvoir être utilisé de manière générale, puisqu'en fonction des valeurs de capacité de soutien utilisées il permet de simuler :

i) la colonisation par *B* d'une aire géographique déjà peuplée par *A*;

ii) la mise en place d'un équilibre démographique entre *A* et *B* ;

iii) l'impossibilité pour *A* de coloniser une aire déjà peuplée par *B*.

Ce modèle est donc idéal pour être utilisé dans la simulation d'un monde hétérogène pour K_A et K_B, dans lequel certaines régions sont plus favorables, en termes de densité, à l'une ou l'autre des populations. Une application possible de ce modèle serait la simulation de la diffusion de l'agriculture à l'échelle mondiale, en tenant compte de la végétation. L'économie de prédation pourrait être considérée comme avantagée par rapport à l'économie de production dans certains

environnements spécifiques comme la forêt tropicale ou le désert, puisque ce sont des environnements dans lesquels le mode de vie chasseur-collecteur a subsisté jusqu'à nos jours, alors qu'il a disparu de la plupart des autres types d'environnement (p. ex. : Jobling *et al.* 2004).

4.3.2 Migrations

4.3.2.1 Migrations intrapopulationnelles

Pour simuler les migrations d'individus entre dèmes voisins et appartenant à la même population (*A* ou *B*), nous avons utilisé un modèle qui considère un taux d'émigration *m* constant. Ce taux d'émigration est indépendant de la taille de la population concernée. Le nombre effectif *E* d'émigrants dans un dème *i* au temps *t* peut s'écrire :

$$E_{i,t} = mN_{i,t} \qquad \text{(Eq. 4.13)}$$

où $N_{i,t}$ est la densité dans la cellule *i* au temps *t*.

Lorsque la population est à l'équilibre, le nombre total d'émigrants *E* vers les 4 cellules voisines est alors égal à $K_i m$. Il s'agit du paramètre "*Nm*" dont les effets sur les généalogies de gènes a été abondamment décrit dans le chapitre 3.

Les émigrants sont ensuite répartis dans les cellules voisines, en fonction de leur friction. La friction d'une cellule est un indice représentant la difficulté pour un individu de se mouvoir à l'intérieur de celle-ci. Chaque cellule possède son propre coefficient de friction *F*, calculé sur la base de données environnementales (Ray 2003: chapitre 3). Une probabilité directionnelle (D_j) de migration vers chacune des quatre cellules est ensuite calculée, en fonction de leurs frictions relatives, comme :

$$D_j = \frac{1-F_j}{\sum_{j=1}^{4} 1-F_j} , \qquad \text{(Eq. 4.14)}$$

où F_j est la friction de la cellule voisine *j*. Ainsi, plus la friction relative d'une cellule est élevée et moins il sera probable qu'un migrant y pénètre. D_j étant une probabilité relative, il est évident que :

$$\sum_{j=1}^{4} D_j = 1 . \qquad \text{(Eq. 4.15)}$$

Le nombre d'émigrants dans chacune des quatre directions peut donc s'écrire comme

$$E_{ij,t} = mN_{i,t} \cdot D_j \qquad \text{(Eq. 4.16)}$$

Lorsque la friction n'est pas prise en compte ou lorsque l'on se trouve dans un milieu homogène, alors D_j est égal à 0.25 dans toutes les directions. Le nombre d'émigrants dans chacune des quatre directions devient alors :

$$E_{ij,t} = \frac{mN_{i,t}}{4} \qquad \text{(Eq. 4.17)}$$

4.3.2.2 Migrations interpopulationnelles ou hybridation

Le flux génique entre les deux populations *A* et *B* peut être modélisé par des migrations d'une population vers l'autre. Ces migrations interpopulationnelles correspondent aux effets des mariages mixtes[1] entre individus appartenant à chacune des deux populations. Il faut noter que, dans le cas de la simulation du Néolithique européen, les migrations interpopulationnelles représentent également l'adoption de l'agriculture par les chasseurs-collecteurs, un phénomène aussi appelé "acculturation" (Ammerman et Cavalli-Sforza 1984). Les conséquences génétiques d'un événement d'acculturation sont identiques à celles d'un mariage mixte. Dans les deux cas, un individu appartenant à une des populations va avoir au moins un ancêtre dans l'autre population à la génération précédente.

La probabilité d'un mariage mixte dépend étroitement des densités respectives des deux populations dans une cellule. Le nombre total de mariages possibles dans une cellule est égal à $\frac{(N_A + N_B)^2}{4}$. Le nombre de mariages mixtes possibles dans la même cellule est égal à $2\frac{N_A N_B}{4}$. La probabilité d'un mariage mixte parmi tous les mariages possibles est donc égale à $P(AB) = \frac{2N_A N_B}{(N_A + N_B)^2}$. Chaque individu *A* a une probabilité *P(AB)* d'effectuer un mariage mixte, en admettant qu'il ne fait pas de distinction entre congénères et individus de l'autre population. Il est cependant fort probable qu'un individu *A* choisira favorablement un individu de la même population par rapport à un individu *B*. Pour refléter ce choix préférentiel, nous introduisons une variable γ qui représente la probabilité d'un mariage mixte en cas de rencontre. γ peut également être décrite comme la probabilité d'hybridation entre les deux populations. Si γ vaut 1, cela signifie que les mariages ont lieu indépendamment des populations auxquelles appartiennent les conjoints. En revanche si γ vaut 0, cela veut dire que les membres de chaque population ne se marieront qu'entre eux. Le nombre S_{AB} de migrations interpopulationnelles de *A* vers *B*, pendant une génération, est fonction de γ, mais également de la probabilité d'effectuer un mariage avec un individu de l'autre population. Nous utiliserons dorénavant le terme "**hybridation**" à la place de migration interpopulationnelle. Le nombre absolu S_{AB} d'événements d'hybridation par génération peut donc s'écrire comme :

$$S_{AB} = \gamma N_A \frac{2N_A N_B}{(N_A + N_B)^2} \qquad \text{(Eq. 4.18)}$$

Ce modèle d'hybridation est très proche de celui utilisé par Barbujani *et al.* (1995), ces mêmes auteurs mentionnant que les résultats obtenus avec leur modèle sont eux-mêmes très proches de

---

[1] Le terme "**mariage**" ne correspond évidemment pas à sa définition religieuse ou civile, mais est utilisé pour définir le choix d'un partenaire sexuel.

ceux obtenus par Rendine *et al.* (1986) avec un modèle plus simple ($S = \gamma N_A N_B$, voir section 4.2). Le raisonnement sous-jacent aux modèles développés dans le cadre de ces deux études n'est pas très différent du nôtre. Leur point commun est l'incorporation d'individus *A* dans la population *B* en fonction des densités des deux populations présentes. C'est par des détails à l'intérieur de ces fonctions que se fait la différence entre les modèles (voir section 4.2). En revanche, Rendine *et al.* (1986) et Barbujani *et al.* (1995) accordent une signification différente à la variable γ, qui représente le taux d'acculturation, puisqu'elle est utilisée uniquement dans le cadre de la transition néolithique.

Contrairement aux études préalables, notre modèle permet un flux génique allant dans les deux sens, soit de *A* vers *B* mais aussi de *B* vers *A*. Pour calculer le nombre S_{BA} d'individus *B* qui passe dans la population *A* pendant une génération, il suffit d'inverser les indices de l'équation 4.18.

L'ajout de possibilités d'hybridation au modèle de compétition peut passablement modifier l'état final du système, par rapport à celui attendu sous le simple modèle de compétition. La résolution analytique d'un tel système non linéaire, dont le nombre de variables est important ($N_{A,0}$, $N_{B,0}$, K_A, K_B, r_A et r_B, γ_{AB} et γ_{BA}) sort des limites de nos compétences. En revanche, la compréhension du système est possible grâce à des simulations. Avant d'utiliser ce modèle dans un monde constitué de cellules hétérogènes, il est donc important de choisir les paramètres de manière adéquate, afin qu'ils correspondent à l'état final désiré. Nous ne sommes cependant pas confronté à ce problème dans le cadre de cette thèse, puisque nous n'utilisons qu'un cas particulier du modèle, pour lequel le monde est homogène et K_B est toujours plus grand que K_A. Cela implique que la population *A* disparaît dans toutes les situations.

4.3.3 Cycle démographique complet

Le modèle démographique incluant à la fois la compétition intrapopulationnelle et interpopulationnelle, ainsi que les migrations intrapopulationnelles et l'hybridation, a été implémenté dans le logiciel SPLATCHE pour qu'à chaque génération, et dans chaque cellule, un cycle démographique se passe selon les phases ci-dessous. Les détails de cette implémentation sont présentés dans l'ANNEXE 4.

Phase 1: Hybridation interpopulationnelle

Les hybrides issus de parents appartenant à des populations différentes sont supprimés de la population *A* et ajoutés à la population *B*, et *vice versa*, selon :

$$N'_{A,t} = N_{A,t} + \frac{2N_{A,t}N_{B,t}}{(N_{A,t} + N_{B,t})^2}\left(-\gamma_{AB}N_{A,t} + \gamma_{BA}N_{B,t}\right) \qquad \text{(Eq. 4.19)}$$

et

$$N'_{B,t} = N_{B,t} + \frac{2N_{A,t}N_{B,t}}{(N_{A,t} + N_{B,t})^2}\left(+\gamma_{AB}N_{A,t} - \gamma_{BA}N_{B,t}\right)$$

Phase 2: Régulation démographique (sélection juvénile)

De nouvelles densités *N"* sont calculées en fonction des naissances et des décès à l'intérieur même de chacune des deux populations et peuvent s'écrire comme :

$$N"_{A,t} = N'_{A,t}\left(1 + \frac{r_A(K_A - N'_{A,t} - \alpha_{AB}N'_{B,t})}{K_A}\right) \qquad \text{(Eq. 4.20)}$$

et

$$N"_{B,t} = N'_{B,t}\left(1 + \frac{r_B(K_B - N'_{B,t} - \alpha_{BA}N'_{A,t})}{K_B}\right)$$

Phase 3: Migrations intrapopulationnelles (post sélection)

Les deux populations échangent un certain nombre de migrants avec les cellules voisines appartenant à la même population. Les densités des deux populations sont alors mises à jour en fonction de ces migrations intrapopulationnelles :

$$N_{A,t+1} = N"_{A,t} - mN"_{A,t} + I_{A,t} \qquad \text{(Eq. 4.21)}$$

et

$$N_{B,t+1} = N"_{B,t} - mN"_{B,t} + I_{B,t}$$

où le premier terme à droite de l'égalité représente la densité de la population après régulation démographique, et les deux derniers termes respectivement le nombre d'émigrants et le nombre d'immigrants intrapopulationnels.

4.3.3.1 Ordre des phases de régulation et de migration

Nous avons décidé de procéder en premier lieu à la phase de régulation démographique, avant la phase de migration intrapopulationnelle. Ceci permet de modéliser de la sélection ou mortalité infantile et des migrations adultes. Cependant, il serait tout aussi cohérent d'inverser ces deux phases, mais cela n'aurait qu'un impact négligeable sur les résultats (non montrés). Le seul effet d'une telle inversion est de légèrement ralentir l'expansion démographique de la population *B*. En effet, lorsque les migrations intrapopulationnelles se font avant la régulation, le nombre de migrants intrapopulationnels est par conséquent légèrement plus faible. Ces différences peuvent être considérées comme négligeables.

4.3.3.2 Simulation typique de l'évolution de deux populations dans la même aire

La Figure 4.8 présente une simulation typique de l'expansion d'une population dans une aire déjà peuplée. Il s'agit de la diffusion d'une population *A* au temps 0 depuis le centre d'un monde carré virtuel, constitué de 50 cellules de côté. Après 500 générations, 100 individus sont tirés de la population *A* pour constituer la population *B* dans la cellule centrale. Il s'ensuit une expansion de la population *B*, couplée à une extinction progressive de la population *A* à cause de la compétition

interpopulationnelle. Il est intéressant de noter qu'un front d'expansion circulaire d'avancée des individus *B* s'établit autour du centre de l'expansion. Puis, une seconde vague circulaire d'extinction des individus *A* est observée légèrement en retrait de la première. C'est pendant la période située entre ces deux vagues d'expansion que se trouve la période de cohabitation entre les deux populations (la bande noire dans la Figure 4.8). C'est uniquement pendant cette période de cohabitation que de l'hybridation peut avoir lieu. La Figure 4.9 illustre, pour la même simulation, l'évolution des densités dans le dème *A*, ainsi que dans le dème *B*, de la cellule centrale.

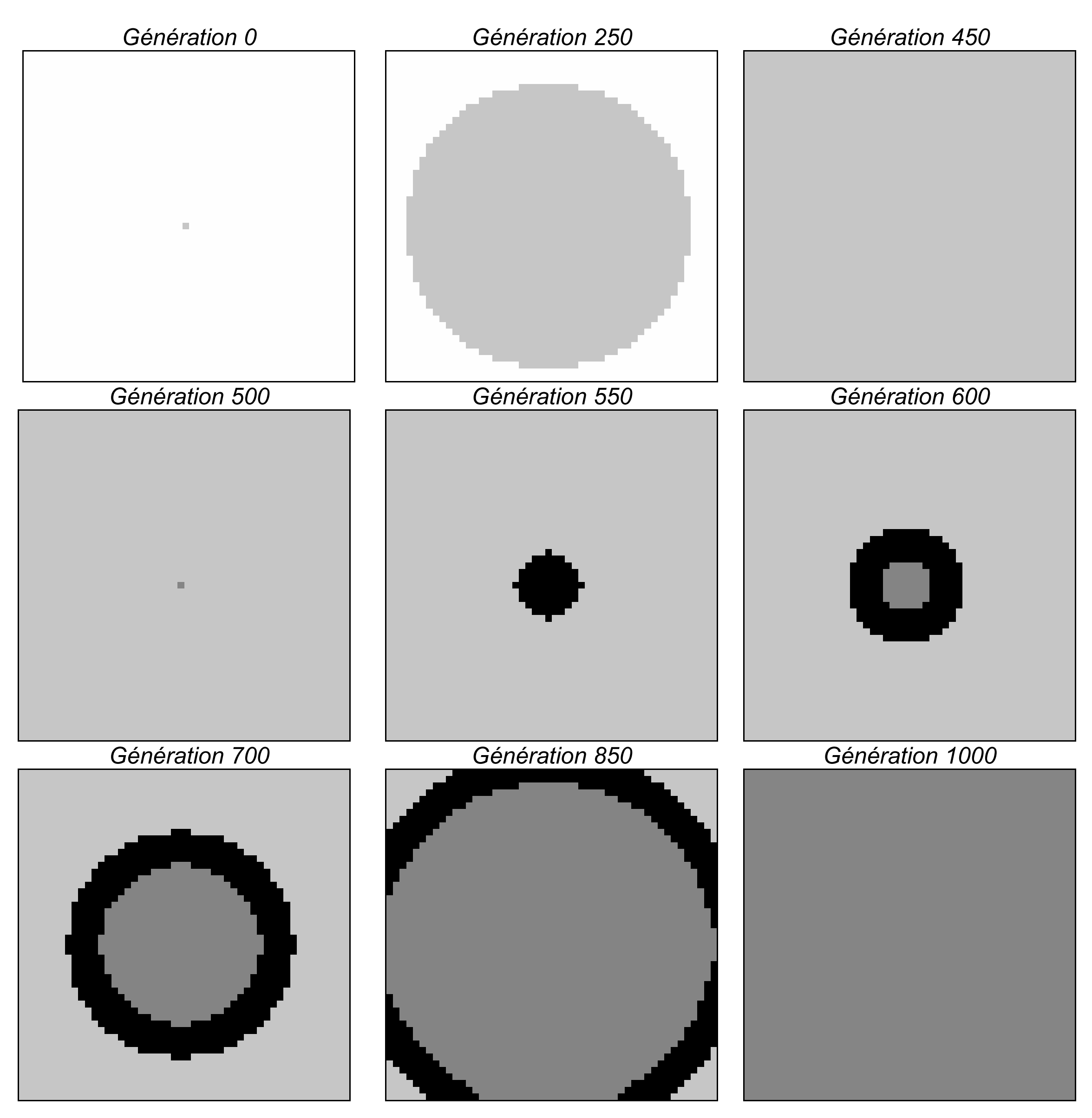


Figure 4.8 Simulation, pendant 1000 générations, de l'expansion de la population *A* au temps 0, puis de celle de *B* au temps 500 depuis le même endroit. L'aire virtuelle est constituée de 50x50 dèmes, avec m = 0.2. Blanc = cellule inoccupée, Gris clair = cellule occupée uniquement par *A*, Gris foncé = cellule occupée uniquement par *B*, Noir = cellule occupée par les deux populations.

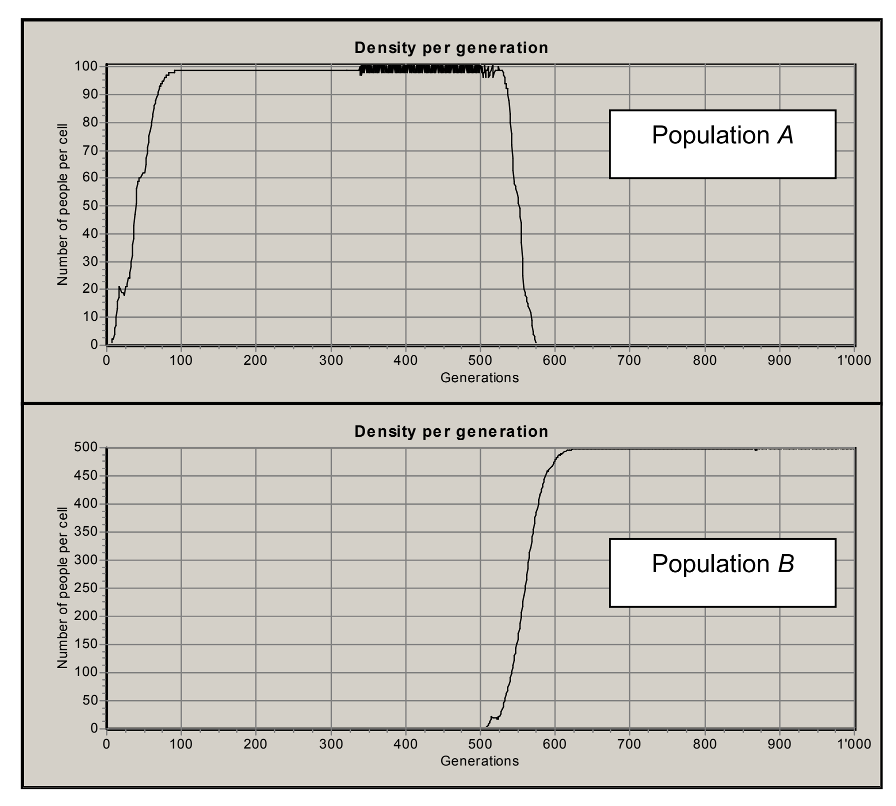


Figure 4.9 Evolution des densités des populations *A* et *B* dans le dème situé au centre d'une aire carrée, constituée de 50x50 cellules, ainsi que dans un dème situé en périphérie. Ces densités ont été stockées en mémoire virtuelle pendant 1000 générations, après l'expansion de la population *A* au temps 0 depuis la cellule centrale, et celle de *B* au temps 500 depuis le même endroit. Avec K_A = 100, K_B=500, r_A = 0.1 et r_B = 0.1. Aucun échange génétique n'a lieu dans ce cas.

4.4 Avantages de l'approche proposée

L'article le plus novateur concernant la modélisation des interactions entre deux populations humaines est celui de Rendine *et al.* (1986). En, effet, même si la méthodologie présentée dans cette publication souffre de certaines lacunes (voir section 4.2), elle a indubitablement inspiré les publications ultérieures, notamment celle de Barbujani *et al.* (1995) qui décrit les simulations les plus réalistes effectuées à ce jour. Notre approche s'inspire de ces modèles, ainsi que d'autres (Aoki 1996 ; Aoki *et al.* 1996), mais nous avons cependant développé notre propre méthodologie pour plusieurs raisons :

1. Nous voulions avoir un contrôle complet sur le modèle utilisé et bien en cerner tous les aspects, ce qui n'est pas toujours possible à partir de la simple description faite dans les publications. Il existe, en effet, de nombreuses incohérences dans les modèles présentés, notamment dans celui de Rendine *et al.* (1986) qui sert de référence aux autres publications (voir section 4.2).

2. La signification du paramètre γ, qui représente le taux d'acculturation dans les modèles antérieurs au nôtre, n'est pas toujours très claire et ne nous satisfaisait pas pleinement. Nous avons

donc redéfini γ d'une manière qui permette de l'utiliser de façon plus générale. Selon notre modèle, γ est le paramètre qui permet de réguler les échanges génétiques entre populations.

3. Notre modèle permet une plus grande variabilité du nombre d'hybridations entre la population *A* et la population *B* lors de chaque génération. Dans les autres études, γ est un paramètre fixé et l'amplitude des migrations interpopulationnelles (hybridation) est très restreinte.

4. Aucun des modèles existants ne permet de simuler de la compétition entre les deux populations (voir section 4.2). La disparition de *A* ne se faisait que sous l'effet de l'assimilation, et de ce fait la période de contact entre les deux populations était extrêmement longue, voire infinie, ce qui ne nous paraîssait pas très réaliste.

De plus, l'implémentation de notre modèle démographique dans une version modifiée de SPLATCHE (voir ANNEXE 4) offre également les avantages suivants :

5. Une plus grande souplesse quant à la variation des paramètres. En effet, l'utilisation de la coalescence alliée à la puissance informatique dont nous bénéficions à l'heure actuelle, notamment grâce au cluster de 40 machines du "CMPG"[1], permet d'explorer l'espace des valeurs de paramètres possibles dans un laps de temps raisonnable. Ceci est particulièrement important du fait que la plupart des variables démographiques des populations réelles sont très mal connues (voir section 4.5.2). L'incertitude autour de la pertinence du modèle et surtout des valeurs des paramètres est donc compensée, en partie, par une exploration intensive de l'espace des résultats possibles. De manière générale, l'effet de la variation des différents paramètres sur les résultats est très peu étudié dans les études effectuées antérieurement.

6. La simulation dynamique de la population *A*. Dans les études antérieures, le rôle de la population *A* (chasseurs-collecteurs) est réduit, au mieux, à une présence au début de l'expansion spatiale et démographique de la population *B*. Aucune attention n'est portée sur l'influence de la dynamique spatiale de la population *A* sur la structure génétique.

7. L'influence de l'environnement sur les paramètres démographiques tels que les densités ou les migrations. Le milieu ne joue en effet, pratiquement aucun rôle dans les études préalables.

4.5 Comportement du modèle

Nous avons décidé d'étudier l'influence des différents paramètres de notre modèle sur les données génétiques, dans un cadre géographique et temporel identique à celui utilisé dans le chapitre 3. La réutilisation de ce cadre permet, d'une part, de nous appuyer sur les obervations du chapitre 3 pour tester l'influence des nouveaux paramètres sur les généalogies et sur la diversité moléculaire, et d'autre part, de comparer les observations obtenues après l'expansion d'une population dans une aire préalablement vide avec celles obtenues par la même expansion dans une aire déjà peuplée.

---

[1] Computational and Molecular Population Genetics Laboratory, à l'Université de Berne.

4.5.1 Schéma de simulation

A la diffusion d'une population d'*Homo sapiens sapiens* dans un monde vide, il y a environ 100'000 ans (comme dans la section 3.2), succède une seconde expansion, qui correspond à la transition Néolithique il y a environ 10'000 ans, qui s'est déroulée de façon relativement contemporaine dans différentes régions du monde (voir par exemple Ammerman et Cavalli-Sforza 1984: pp. 13-16 ; Bellwood 2001 ; Jobling *et al.* 2004 et les références qui y sont mentionnées). Evidemment, les dates simulées sont approximatives, mais elles permettent de tester notre modèle dans un cadre réaliste. Nous reviendrons plus loin (page 88) sur l'influence du cadre temporel sur les données génétiques simulées.

Nous avons donc procédé à une série de simulations dans un monde simple, homogène pour *K* et *F* (friction), représenté par une matrice carrée de 2'500 dèmes. L'expansion paléolithique de la population *CC* (pour chasseurs-collecteurs) prend sa source dans le dème central <25 ; 25>. Après 3'600 générations, une seconde expansion (néolithique) se déroule dans la matrice *AG* (pour agriculteurs), soit depuis la même cellule centrale <25 ; 25>, soit depuis une cellule périphérique <5 ; 5>. La population initiale est composée de N_{CC} individus qui apparaissent *in situ*, alors que dans le cas de la seconde expansion, N_{AG} individus sont tirés d'un unique dème de la population *CC* pour créer la nouvelle population *AG* (stratégie 1 : Figure 9.16 dans l'ANNEXE 4). Dans toutes les simulations, la capacité de soutien de la population néolithique est supérieure à celle de la population de chasseurs-collecteurs ($K_{AG} > K_{CC}$), comme il est communément admis (Hassan 1979 ; 1981 ; Rendine *et al.* 1986 ; Langaney *et al.* 1990; Landers 1992 ; Barbujani *et al.* 1995 ; Cavalli-Sforza 1996 ; Pennington 2001 ; Diamond et Bellwood 2003 ; Gallay 2004). Dans ce cas de figure, l'état final du système correspond à la disparition des chasseurs-collecteurs, ce qui correspond à une observation avérée dans la plupart des régions de l'Ancien Monde, et notamment en Europe.

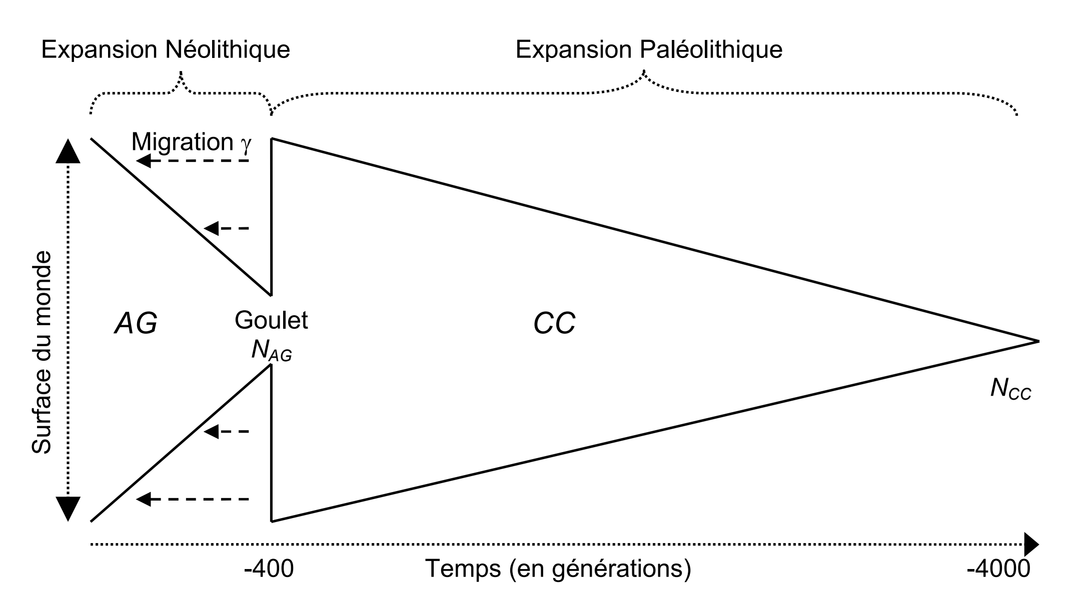


Figure 4.10 Schéma des simulations dans un monde carré homogène pour *K* et *F* (friction) et constitué de 2'500 dèmes.

Dans certains cas, des migrations interpopulationnelles (hybridation) peuvent avoir lieu à partir de la population *CC* vers la population *AG*, à un taux γ (Figure 4.10). Ces migrations représentent à la fois les mariages mixtes entre ces deux populations et l'adoption de l'agriculture par les chasseurs-collecteurs. Les enfants issus de ces deux processus appartiennent à la population *AG* et possèdent au moins un ancêtre dans la population *CC*. Si on ne peut évidemment pas exclure que des échanges génétiques aient eu lieu dans les deux sens, nos simulations ont confirmé que cela n'a quasiment aucune influence sur les résultats lorsque la population *CC* disparaît après quelques générations de contact, comme c'est le cas ici. Par conséquent, nous avons décidé de ne considérer qu'un flux génique allant dans le sens de la population *CC* vers la population *AG*, comme l'ont fait d'autres auteurs avant nous (Rendine *et al.* 1986 ; Barbujani *et al.* 1995 ; Aoki 1996 ; Aoki *et al.* 1996). La Figure 4.11 illustre schématiquement l'évolution des densités des populations *CC* et *AG*, ainsi que les hybridations, selon le cas particulier du modèle que nous utilisons ici, soit $K_{CC} < K_{AG}$ et hybridation uniquement de *CC* vers *AG* ($S_{CC\text{->}AG}$).

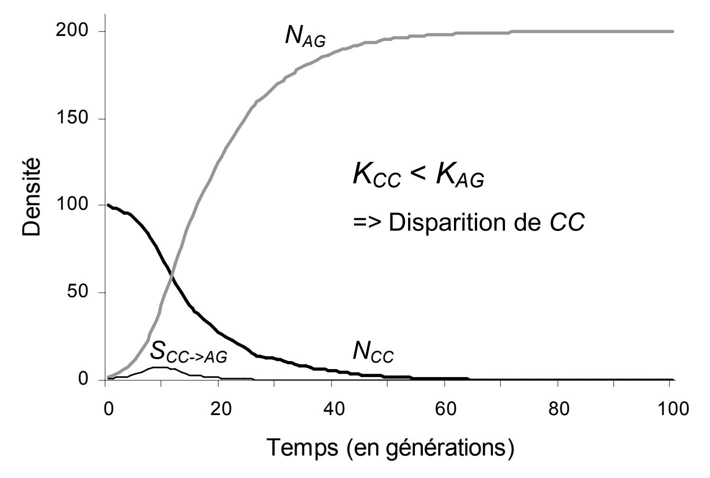


Figure 4.11 Schéma de l'évolution des densités des populations *CC* (en gris) et *AG* (en noir) au cours du temps, ainsi que des hybridations allant uniquement de *CC* vers *AG* (trait fin). $K_{CC} < K_{AG}$.

L'influence du goulet d'étranglement néolithique étant sans aucun doute sous-estimée de par le fait que notre modèle ne tolère qu'une seule coalescence par génération, nous avons modifié ce dernier de manière à ce qu'un maximum de 10 coalescences soient possibles pendant la génération pendant laquelle a lieu le goulet (génération -400). Lorsque la taille du goulet n'est pas explicitement mentionnée, il s'agit alors d'un goulet constitué d'un seul dème contenant 50 gènes efficaces. Nous verrons dans la section 4.5.3, ce que nous pouvons dire de l'influence de la taille de ce goulet.

4.5.2 Estimation des paramètres

Avant de se lancer dans une série de simulations, il importe de définir un intervalle de valeurs réalistes pour chacun des paramètres démographiques du modèle. Cet espace de valeurs de paramètres raisonnables permettra, d'une part, de limiter le nombre de simulations à effectuer et, d'autre part, de pallier l'imprécision qui existe dans les estimations des variables démographiques. Nous allons donc passer en revue les valeurs attribuées dans la littérature aux différents paramètres

démographiques des populations humaines, et essayer de les convertir en valeurs utilisables à l'aide de notre approche. La Table 4.1 résume les estimations des densités et des taux de croissance tirées de la littérature et mentionnées dans le texte ci-dessous.

Densité CC	***Densité AG***	***Facteur***	***Crois. CC***	***Crois. AG***	***Référence***
-	-	-	1.2	-	Birdsell 1957
< 1.0	-	-	-	-	Lee et DeVore 1968a
-	-	-	1.05	-	Mosimann et Martin 1975
0.01-1.0	-	-	< 0.003	-	Hassan 1979
0.03-2.0	-	x27	< 0.003	-	Hassan 1981
0.02-0.03	-	-	-	-	Hewlett *et al.* 1982
0.01-1.0	3-288	x100	-	0.15-0.8	Ammerman et Cavalli-Sforza 1984, théorie
0.02	0.4	x20	0.25	0.5	Ammerman et Cavalli-Sforza 1984, simulations
0.28	-	-	-	-	Weiss 1984
0.04	1.07	x28	0.25	0.5	Rendine *et al.* 1986
-	-	-	0.6	-	Winterhalder *et al.* 1988
-	-	-	0.87	-	Belovsky 1988
0.1-1.0	-	-	0.006	0.015	Landers 1992
0.1	0.75	x7.5	0.25	0.5	Calafell et Bertranpetit 1993
-	-	-	0.2-0.3	-	Cavalli-Sforza *et al.* 1994
0.1	-	-	0.75	-	Young et Bettinger 1995
0.04	0.9	x22	-	0.5	Barbujani *et al.* 1995
-	-	x10	-	-	Cavalli-Sforza 1996
-	0.15	-	-	0.8	Fix 1997
0.015-0.2	-	-	0.3-0.9	-	Steele *et al.* 1998
-	-	-	0.12	-	Anderson et Gillam 2000
0.026	-	-	-	-	Bocquet-Appel et Demars 2000a
0.02-100	3-300	x3-x150	-	-	Zvelebil 2000
0.02-100	1-70	x0.7-x50	-	-	Zvelebil 2000
0.0024	-	-	0.9	-	Pennington 2001
0.04-0.08	-	-	0.5	-	Alroy 2001
0.03	-	-	0.07	-	Eswaran 2002
0.01-0.35	-	-	-	-	Binford 2001; Ray 2003
-	-	x10	-	-	Biraben 2003

Table 4.1 Estimations des densités (par km^2) et des taux de croissance (par génération) des populations de chasseurs-collecteurs (*CC*) et des populations néolithiques (*AG*).

La plupart des informations dont nous disposons sur la démographie des populations de chasseurs-collecteurs paléolithiques et mésolithiques sont tirées de comparaisons faites sur la base d'observations de populations de chasseurs-collecteurs contemporains[1]. Cependant, rien ne nous permet d'affirmer que les ethnies de chasseurs-collecteurs actuelles ont une démographie comparable aux populations du passé. En effet, alors que ces dernières exploitaient pratiquement tous les types de terrains (Roebroeks 2001), les chasseurs-collecteurs contemporains ne subsistent que par petits groupes et sont confinés dans des zones le plus souvent défavorables à l'agriculture (forêt tropicale, désert), qui sont très éloignées des zones optimum qu'ils exploitaient jadis (Spielmann et Eder 1994 ; Pennington 2001 :p. 311). De plus, les ethnies contemporaines sont

[1] Voir Pennington (2001) pour une revue de la démographie des ethnies actuelles de chasseurs-collecteurs.

soumises à la pression exercée par l'agriculture et les communautés industrialisées voisines, avec qui elles ont des contacts plus ou moins intensifs (Spielmann et Eder 1994 ; Blurton Jones *et al.* 2002). Ces voisins modernes réduisent, d'une part, l'aire d'influence des ethnies de chasseurs-collecteurs et d'autre part, leur transmettent de nouvelles maladies infectieuses (Dunn 1968 ; Lee et DeVore 1968a ; Landers 1992). D'un autre côté, les données archéologiques ne permettent pas d'estimation précise des densités de populations humaines avant leur sédentarisation, c'est pourquoi les données ethnographiques restent tout de même les meilleurs estimateurs, malgré leurs défauts (voir Ray 2003: p. 32 pour une plus ample discussion à ce sujet).

Comme nous l'avons vu dans la section 4.3, le modèle démographique utilisé dans ce travail pour simuler la croissance d'une population est une courbe logistique qui requiert deux paramètres : le paramètre *r,* qui permet de réguler la vitesse de croissance de la population; le paramètre *K*, la capacité de soutien, qui peut être mise en relation avec la densité.

4.5.2.1 Taux de croissance[1]

Les estimations du taux de croissance sont de trois types: 1° comparaison avec des ethnies de chasseurs-collecteurs ou d'agriculteurs contemporains ; 2° estimations sur la base du peuplement du continent américain ; 3° croissance à long terme pour passer de quelques milliers d'individus, il y a 100'000 ans, à quelques milliards aujourd'hui. Les estimations de la croissance globale de l'espèce humaine sur une longue période ne sont cependant pas comparables avec le taux de croissance d'une population sur une courte durée. Sur une longue durée, une population passe en effet plus de temps à capacité de soutien avec une croissance nulle, qu'en croissance. Comme nous le verrons ci-dessous, les estimations sur le long terme sont généralement beaucoup plus faibles que celles effectuées sur un petit nombre de générations, sans pour autant être incompatibles.

- Chasseurs-collecteurs (r_{CC})

Les estimations faites par comparaison avec des communautés contemporaines de chasseurs-collecteurs ont révélé un taux de croissance *r* pouvant aller jusqu'à 80% par génération dans ces populations (Pennington 2001). Un taux de croissance aussi important est le maximum généralement admis pour l'espèce humaine (Ammerman et Cavalli-Sforza 1984; Young et Bettinger 1995) et serait caractéristique des populations qui colonisent des territoires inoccupés, disposant de ressources abondantes. Un tel taux de croissance ne serait donc possible que pendant une courte période, au moment de l'arrivée des premiers colons dans une zone déserte. C'est exactement ce qui se passe avec le modèle de croissance logistique, puisque la valeur de *r* n'est atteinte que pendant les premières générations (Figure 4.3), la croissance démographique diminuant par la suite sous l'effet de la limitation des ressources locales. C'est donc seulement au front de la vague d'expansion qu'une croissance instantanée égale à la valeur de *r* peut avoir lieu. Steele *et al.* (1998) suggèrent cependant que les estimations tirées de comparaisons ethnographiques sont très

---

[1] Toutes les valeurs de taux de croissance mentionnées dans ce travail sont données en générations.

éloignées de la valeur maximum du taux de croissance, car les populations de chasseurs-collecteurs actuelles sont proches de leur capacité de soutien.

Les nombreuses estimations qui ont été faites sur la base du peuplement du continent américain situent généralement le taux de croissance entre 12% et 90% (Belovsky 1988 ; Winterhalder *et al.* 1988 ; Steele *et al.* 1998 ; Anderson et Gillam 2000 ; Alroy 2001), même si des estimations antérieures font état de valeurs inférieures (3% : Hassan 1981) ou supérieures (>100% : Birdsell 1957 ; Mosimann et Martin 1975).

Les estimations du taux de croissance à long terme de la population humaine sont généralement beaucoup plus basses (0.3%-0.6%) puisqu'elle font l'hypothèse que l'humanité est passée, de façon exponentielle, de quelques milliers d'individus, il y a 100'000 ans, à environ 5 à 10 millions à la fin du paléolithique (Coale 1974; Hassan 1981 ; Landers 1992). Ces valeurs ne sont pas du tout incompatibles avec celles tirées d'observations ethnographiques ou basées sur le peuplement de l'Amérique, puisque, comme nous l'avons déjà mentionné, sur une longue durée une population passe plus de temps à l'équilibre qu'en croissance. Ainsi le taux de croissance à long terme est faible, alors que pendant la (ou les) période(s) de croissance, ce taux est beaucoup plus important.

Les valeurs ponctuelles généralement retenues ou mentionnées lors de simulations sont de l'ordre de 20% à 30% (Rendine *et al.* 1986; Calafell et Bertranpetit 1993 ; Cavalli-Sforza *et al.* 1994).

- Néolithique (r_{AG})

Il est généralement admis que le taux de croissance des agriculteurs est supérieur à celui des chasseurs-collecteurs (Zvelebil et Zvelebil 1988; Jackes *et al.* 1997), même si ce changement s'est peut-être effectué lentement, notamment par un accroissement du taux de natalité plus important que du taux de mortalité (Coale 1974). Rendine *et al.* (1986), Calafell et Bertranpetit (1993) ainsi que Barbujani *et al.* (1995) ont, par exemple, utilisé un taux de croissance de 50% pour la population néolithique dans leurs simulations, qui est une estimation grossière faite par Ammerman et Cavalli-Sforza (1984: p. 75) sur la base de données archéologiques. Fix, en 1997, a utilisé un taux de 80% dans d'autres simulations et, à titre de référence, le taux de croissance mondiale pour les 50 dernières années est de 60% (Pennington 2001: p. 171). Il existe cependant des estimations beaucoup plus faibles, de l'ordre de 1.5% (Coale 1974; Landers 1992), mais il s'agit à nouveau d'estimations sur le long terme.

4.5.2.2 Densités de population

La plupart des estimations de densité de populations chasseurs-collecteurs ont été faites à l'aide de données ethnographiques, en observant les populations actuelles. Ce sont cependant des valeurs observées localement et dont il est difficile de faire une extrapolation moyenne pour l'Europe paléolithique, d'autant plus que les estimations ethnographiques semblent être assez optimistes pour évaluer les densités préhistoriques (Bocquet-Appel et Demars 2000a).

- Chasseurs-collecteurs (K_{CC})

D'après les comparaisons ethnographiques, les densités des populations de chasseurs-collecteurs contemporains vont de 0.02 à 100 individus par km^2 (Zvelebil 2000), par exemple entre 0.02 et 0.03 chez les pygmées Aka (Hewlett *et al.* 1982). Nous laissons volontairement de côté la valeur de 100 individus par km^2 qui ne peut en aucun cas être prise comme moyenne des chasseurs-collecteurs pour l'Europe, comme le montre le simple calcul suivant : la superficie du continent européen est de 23'594'000 km^2 (Source UNEP[1]) et la population maximum de chasseurs-collecteurs dans le monde à la fin du paléolithique est estimée entre 5 et 10 millions (Lee et DeVore 1968b ; Hassan 1981 ; Landers 1992). Même en considérant que la totalité de la population mondiale (10 millions d'individus) se trouvait en Europe à la fin du paléolithique (ce qui est faux !), on obtient par simple calcul une densité moyenne de 2.36 individus par km^2. Ce chiffre étant largement surestimé, il nous parait raisonnable de ne pas considérer de densités moyennes de chasseurs-collecteurs supérieures à 1.0 individus par km^2.

Les estimations faites par Bindford (2001) dans son magistral ouvrage "*Constructing frames of references*" et reprises par Ray (2003) font état de densités de chasseurs-collecteurs allant de 0.005 dans les déserts à 0.35 dans les forêts tropicales de montagne. Nous écartons volontairement la densité minimum estimée par Binford pour le désert (0.005), car il s'agit d'une situation extrême. En effet, le peuplement du désert est très hétérogène et constitue un cas particulier qu'il est difficile de modéliser.

La plupart des estimations se situent globalement entre 0.01 et 0.3 individus par km^2. Rendine *et al.* (1986) et Barbujani *et al.* (1995) ont, par exemple, utilisé des valeurs moyennes de 0.04 dans leurs simulations, alors que Calafell et Bertranpetit (1993) ont estimé une densité de 0.1 individus par km^2 pour les populations mésolithiques de la péninsule ibérique.

- Néolithique (K_{AG})

Les densités des populations agropastorales actuelles vont de 3 individus par km^2 au Laos ou au Zimbabwe, à 300 en Nouvelle Guinée (Zvelebil 2000). Il s'agit à nouveau d'estimations locales, difficilement applicables à l'ensemble du continent européen. A titre de référence, au 15ème siècle, l'Europe était peuplée, en moyenne, par 1 à 70 individus par km^2 (Zvelebil 2000). Pour de nombreux auteurs, les techniques agricoles ont permis d'atteindre des densités humaines beaucoup plus élevées au Néolithique (Hassan 1979 ; 1981 ; Langaney *et al.* 1990 ; Landers 1992 ; Langaney *et al.* 1992 ; Pennington 2001 ; Cavalli-Sforza et Feldman 2003 ; Gallay 2004) d'une part, parce qu'une plus grande quantité de nourriture a pu être produite sur une surface équivalente et d'autre part, parce que le surplus de nourriture a pu être stocké grâce au mode de vie sédentaire adopté par les populations agropastorales (Diamond et Bellwood 2003). Cependant, pour d'autres auteurs (Zvelebil et Zvelebil 1988 ; Fix 1996 ; Zvelebil 2000), la différence entre les densités des communautés prédatrices et productrices n'est pas si grande. Dans certaines régions, comme la côte atlantique ou la Scandinavie, les densités des populations de la fin du mésolithique sont en tout cas aussi

[1] http://www.unep.org/

importantes que celles des premiers agriculteurs de la même région (Jackes *et al.* 1997 ; Arias 1999). Le modèle de croissance logistique tient compte de cette objection puisque – bien qu'à terme les capacités de soutien des néolithiques soient plus importantes que celles des chasseurs-collecteurs – il faut plusieurs générations pour que la densité effective des agriculteurs atteigne, puis dépasse, celle des chasseurs-collecteurs.

Les estimations des densités néolithiques se situent généralement entre 7.5 fois et 50 fois celle des populations de chasseurs-collecteurs (Rendine *et al.* 1986 ; Calafell et Bertranpetit 1993 ; Barbujani *et al.* 1995).

4.5.2.3 Migrations intrapopulationnelles (*m*)

Ayant développé notre propre modèle de migration et ayant montré que c'est la combinaison du produit de la densité d'une population et du taux de migration (*Nm*) qui est importante et non le taux de migration lui-même (section 3.2.1), nous avons fait varier ce taux de façon à ce que la palette des *Nm* simulés soit la plus large possible. Un *Nm* supérieur à 1'000 ne changeant plus rien à la structure génétique observée (section 3.2.1), nous n'utilisons que des taux de migrations *m* variant de 0.04 à 0.2. Cet intervalle permet de faire varier *Nm* de 2 à 1'000, qui sont des valeurs suffisamment extrêmes pour avoir une bonne représentativité des résultats.

4.5.2.4 Hybridation interpopulationnelle (γ)

Notre modèle d'hybridation étant spécifique à cette étude, nous ne pouvons pas comparer directement les "taux d'acculturation" utilisés ailleurs (Rendine *et al.* 1986 ; Calafell et Bertranpetit 1993 ; Barbujani *et al.* 1995) au paramètre γ utilisé ici. Nous faisons varier γ entre 0 (pas d'hybridation) et 1 (tous les individus ont la même probabilité de se marier, quelle que soit leur population d'origine).

4.5.2.5 Temps de cohabitation

Toujours dans le but de calibrer notre modèle, il est nécessaire de connaître quels sont les intervalles réalistes du temps de cohabitation entre chasseurs-collecteurs et néolithiques dans une aire donnée, qui correspond à un dème dans nos simulations. La Table 4.2 recense les périodes de temps entre l'arrivée des premiers éléments néolithiques et la disparition des derniers éléments mésolithiques dans différentes régions d'Europe. Ces temps de chevauchement sont difficilement comparables entre eux, car la taille des régions étudiées varie considérablement. Ils permettent cependant de se faire une idée de leur ordre de grandeur. Excepté la zone 1 (Figure 4.12), constituée de l'Anatolie, de la Turquie asiatique et de Chypre, où la période d'installation du néolithique est plus longue (il s'agit de la mise en place primaire des éléments constituant le néolithique, Mazurié de Keroualin 2001), il faut au maximum 1'100 ans pour avoir un changement total d'économie dans chacune de ces zones (Table 4.2). La côte atlantique est constituée des régions dans lesquelles la cohabitation entre communautés néolithiques et mésolithiques a été la plus longue, notamment à cause des fortes densités mésolithiques, et pourtant la cohabitation n'y a jamais excédé 1'000 ans (Arias 1999). Par ailleurs, P.-Y. Nicod cite des périodes de

"*quelques siècles au maximum*" pour avoir un changement total d'économie dans une zone donnée (communication personnelle) et "*il paraît difficile d'admettre une longue persistance des sociétés de chasseurs parallèlement au développement des communautés agricoles*" (Gallay 1994).

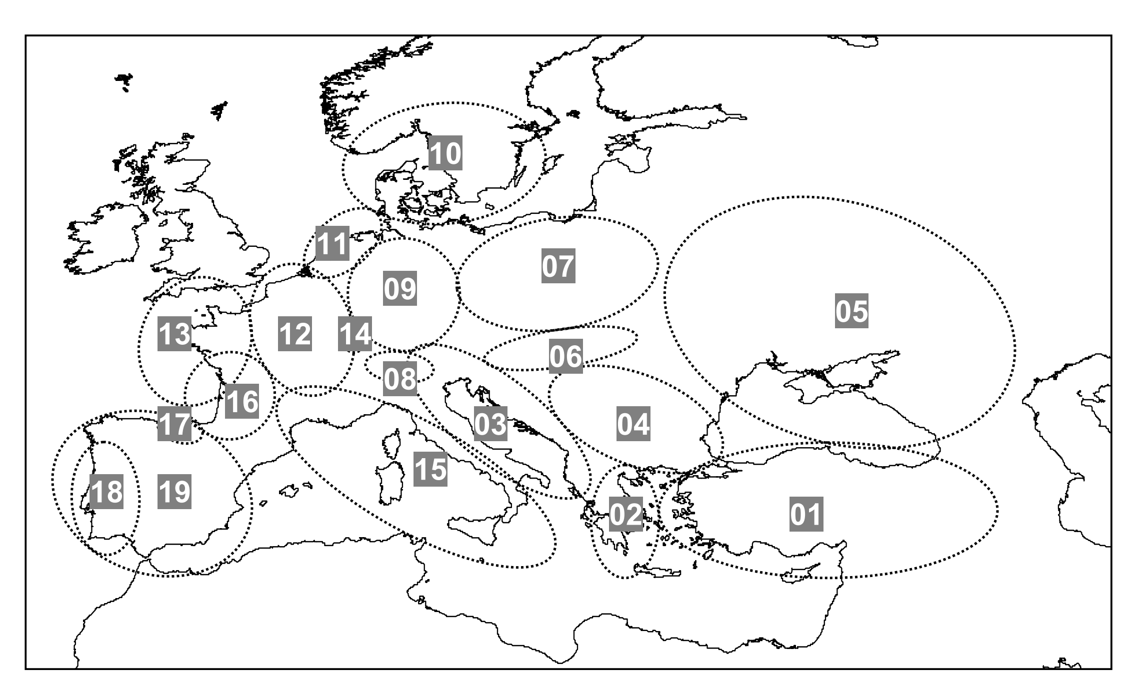


Figure 4.12 Figure illustrant approximativement les régions mentionnées dans la Table 4.2.

Zone Géographique *(le numéro correspond à la Figure 4.11)*	***Période maximum de cohabitation*** *(en années)*	***Référence***
1. Anatolie	2'700	Mazurié de Keroualin 2001
2. Grèce	1'000	Mazurié de Keroualin 2001
3. Adriatique	1'000	Mazurié de Keroualin 2001
4. Balkans	900	Mazurié de Keroualin 2001
5. Nord de la Mer noire	1'000	Mazurié de Keroualin 2001
6. Carpates	1'100	Mazurié de Keroualin 2001
7. Région Elbe-Vistule	1'000	Mazurié de Keroualin 2001
8. Alpes du Nord et Jura	300-700	Gallay 1994
9. Région Elbe-Rhin	500	Mazurié de Keroualin 2001
10. Sud de la Scandinavie	1'000	Arias 1999
11. Mer du nord (Pays-Bas/Allemagne)	1'000	Arias 1999
12. France Nord-Est, Benelux, Suisse	700	Mazurié de Keroualin 2001
13. France côte atlantique	400-900	Arias 1999
14. Franche-Compté	200	Jeunesse 1998
15. Méditerranée occidentale	900	Mazurié de Keroualin 2001
16. France Sud-Ouest	1'100	Mazurié de Keroualin 2001
17. Cantabria (Nord-Ouest de l'Espagne)	800	Arias 1999
18. Portugal	400	Arias 1999
19. Péninsule Ibérique	800	Mazurié de Keroualin 2001, Calafell et Bertranpetit 1993

Table 4.2 Période entre l'arrivée des premiers éléments néolithiques et la disparition des derniers éléments mésolithiques dans différentes régions d'Europe.

Les surfaces comprises dans chacune des zones mentionnées étant beaucoup plus grandes que les dèmes du monde virtuel (50 x 50 *km*) dans lequel nous effectuons nos simulations[1], le temps de cohabitation dans un dème est sans aucun doute beaucoup plus court. Dès lors, il paraît raisonnable de prendre 1'000 ans comme période maximum de cohabitation à l'intérieur d'un dème, ce qui correspond à environ 40 générations humaines, en prenant 25 ans comme temps de génération moyen (le temps de génération chez l'Homme étant vraisemblablement légèrement supérieur : Tremblay et Vezina 2000 ; Helgason *et al.* 2003). Nous écarterons donc toutes les combinaisons de paramètres qui donnent des temps de cohabitation plus longs que 40 générations. Ceci concerne notamment tous les cas où les valeurs de K_{AG} et de K_{CC} sont trop proches.

4.5.2.6 Paramètres utilisés

A partir des estimations tirées de la littérature, nous avons défini un intervalle de valeurs de capacités de soutien K et de taux de croissance r à utiliser (Table 4.3). Nous avons choisi de faire varier les paramètres r_{AG} et r_{CC} entre les valeurs 10% et 80%, et nous avons retenu des densité de chasseurs-collecteurs situées entre 0.02 et 0.8 individus par km^2. L'intervalle de valeurs utilisé pour la densité néolithique se situe entre 0.04 et 8 individus par km^2, l'utilisation de densités supérieures à cet intervalle ne changent rien aux conclusions tirées de nos résultats. Il est cependant nécessaire de convertir ces densités (en individus par km^2) en valeurs de capacité de soutien applicables à nos dèmes virtuels, qui sont en nombres de gènes efficaces. Sachant que la taille efficace d'une population (N_e) est approximativement égale à la moitié de sa taille de recensement ($N_e \approx N/2$ [2]) et que la surface d'un dème représente 2'500 *km*, le nombre de chasseurs-collecteurs efficace est compris entre 25 et 1'000, et celui d'agriculteurs entre 50 et 10'000. Les capacités de soutien pour les chasseurs-collecteurs estimées ici sont compatibles avec les estimations de la taille des groupes en connexion, qui varient entre 150 et 1'000 individus (Lee et DeVore 1968a ; Hassan 1981 ; Cavalli-Sforza et King 1986 ; Roebroeks 2001]), répartis en bandes de 25 (Birdsell 1968 ; Landers 1992). Il faut noter que la capacité de soutien K_{AG} pour les néolithiques est habituellement considérée comme étant entre 2 fois et 50 fois plus grande que celle des chasseurs-collecteurs, soit entre 50 et 50'000 individus effectifs dans notre cas. Cependant, des valeurs de K_{AG} supérieur à 10 fois celle de K_{CC} ne changent quasiment plus rien au nombre de migrants échangés entre les deux populations. Nous utiliserons donc 10'000 comme valeur de K_{AG} maximum. Lors de la simulation de systèmes génétiques haploïdes liés au sexe (génôme mitochondrial ou chromosome Y), il faut diviser les valeurs mentionnées ci-dessus par 2, alors qu'il faut les multiplier par deux lors de la simulation de systèmes diploïdes (comme dans la Table 4.3).

---

[1] Des cellules de 50 *km* de côté ont été jugées comme étant de taille adéquate pour représenter des sous-populations de chasseurs-collecteurs (Anderson et Gillam 2000). Le diamètre de leur territoire saisonnier en Europe centrale il y a environ 8'000 ans, a été estimé entre 60 et 80 *km* (Gronenborg 1999). De plus, la distance d'exploration moyenne des pygmées contemporains s'étend d'environ 50 à 90 *km* (Hewlett *et al.* 1982) et leur distance moyenne de mariage à environ 40 *km* (Cavalli-Sforza et Hewlett 1982), ces valeurs étant sans doute supérieures à celles des chasseurs-collecteurs paléolithiques (Ammerman et Cavalli-Sforza 1984).

[2] La taille efficace N_e peut également être définie comme étant égale à $N/3$, mais pour des raisons de simplification nous utilisons $N/2$ dans ce travail.

Paramètre	*Minimum*	*Maximum*
r_{CC}	0.1	0.8
r_{AG}	0.1	0.8
K_{CC}	50	2'000
K_{AG}	100	20'000
m	0.04	0.2
γ	0	1

Table 4.3 Intervalles des paramètres utilisés. *K* est donné en nombre de gènes efficaces portés par un système diploïde, et *r* en générations.

4.5.3 Influence des paramètres sur la diversité moléculaire

Tout comme pour les simulations présentées dans la section 3.2 (Ray *et al.* 2003) et pour simuler des données proches de celles disponibles pour l'ADN mitochondrial, 1'000 simulations d'un échantillon de 30 séquences de 300 paires de base sont effectuées pour chaque scénario démographique simulé. Dans tous les cas, l'échantillon est prélevé dans la cellule centrale <25 ; 25>, afin d'éviter les légers effets de bord[1] observés lorsque $N_{AG}m$ est petit (voir section 3.2). Nous enregistrons la distribution des différences par paires de séquences (distribution "mismatch" : Rogers et Harpending 1992). Le taux de mutation utilisé est égal à 0.001 et permet d'obtenir des distributions "mismatch" du même ordre de grandeur que celles observées pour le génome mitochondrial dans les populations humaines, à l'échelle mondiale, soit avec un mode attendu d'environ 8[2] (Excoffier et Schneider 1999).

Nous avons premièrement procédé à une série de simulations pour lesquelles la population *AG* remplace complètement la population *CC*, sans aucune hybridation ($\gamma = 0$). La Figure 4.13 permet de visualiser les mouvements, en remontant le temps, d'un échantillon de 30 gènes provenant du centre du monde. On peut voir qu'après une première phase de dispersion dans la population *AG*, les lignages sont ramenés vers la source de l'expansion néolithique. Ils passent ensuite par un goulet d'étranglement, avant de subir une seconde phase de dispersion dans la population *CC*. Finalement, les derniers lignages sont ramenés vers le lieu d'origine de l'expansion paléolithique, où se font les dernières coalescences.

[1] Comportement légèrement différent des dèmes qui se trouvent dans les bords de l'aire simulée par rapport à ceux qui se trouvent au centre, dû aux possibilités de migration restreintes (moins de dèmes voisins). Ce phénomènes a été décrit en détails par Ray (2003: p. 149).

[2] Selon le modèle des sites infinis, $8 = \pi = 2t\mu$ où π est le nombre attendu de différences par paire, *t* est le nombre de générations jusqu'à l'ancêtre commun le plus récent (MRCA) et μ le taux de mutation par génération et pour l'ensemble des locus étudiés.

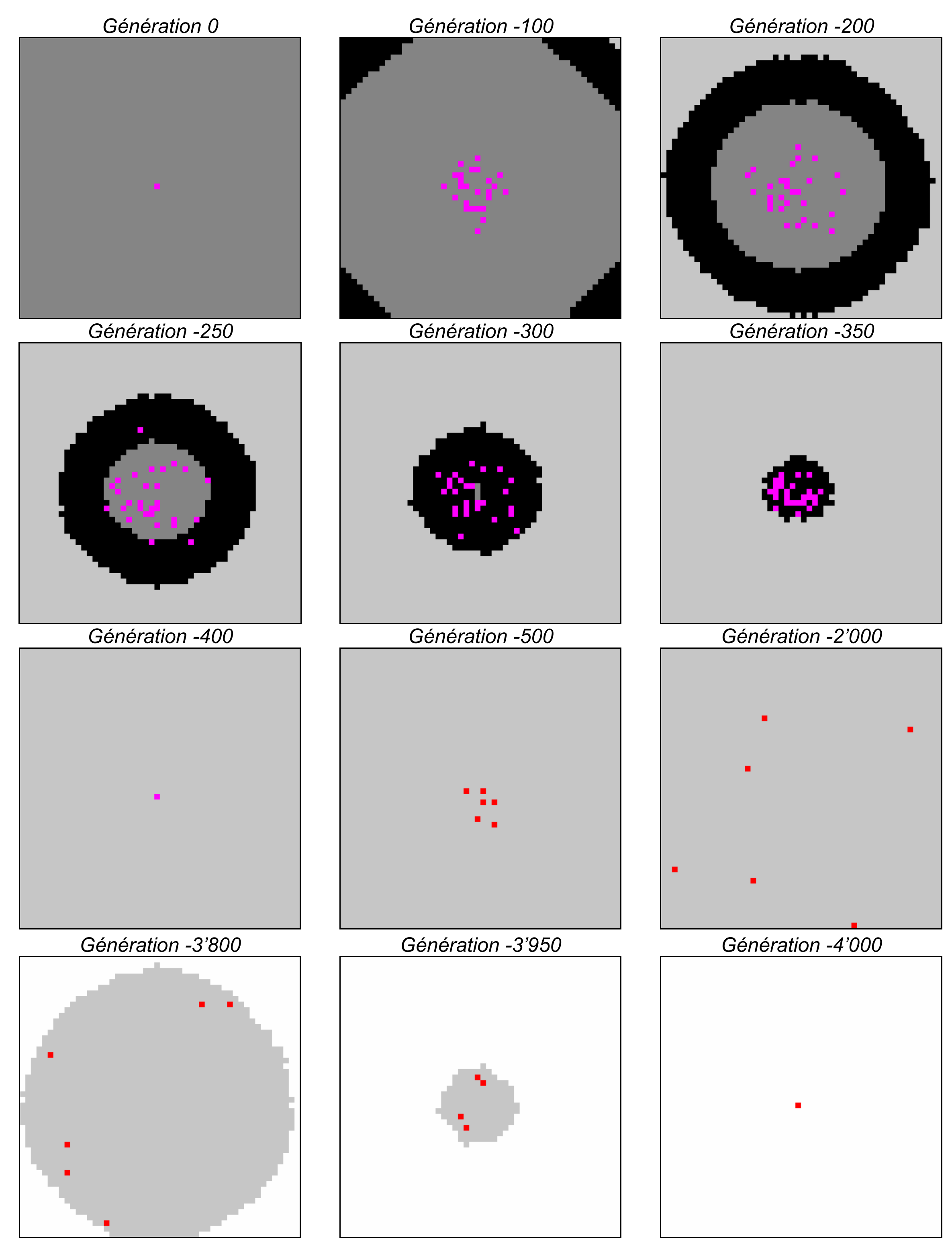


Figure 4.13 Occupation du monde : gris clair = occupation par *CC*, gris foncé = occupation par *AG*, noir = occupation par *CC* et *AG*. Rose = dème de la population *AG* dans lequel se trouve au moins un gène; Rouge = dème de la population *CC* dans lequel se trouve au moins un gène.

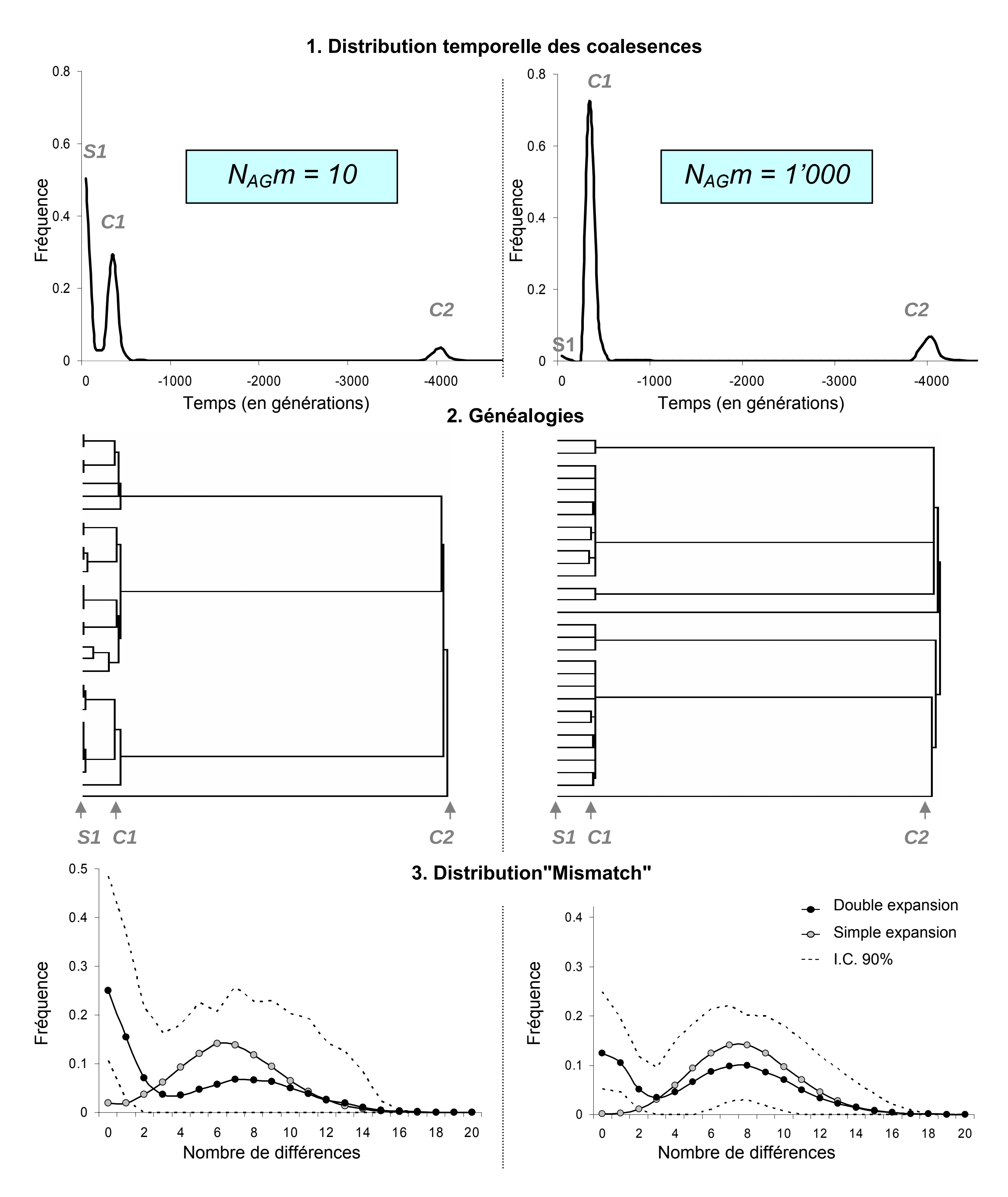


Figure 4.14 Caractéristiques génétiques obtenues après la simulation d'une double expansion démographique et spatiale. $N_{AG}m$ est soit petit (10, colonne de gauche), soit grand (1'000, colonne de droite) : 1. Distribution des coalescences au cours du temps ; 2. Généalogie de gènes échantillonnés; 3. Distribution "mismatch" (cercles noirs) avec I.C. à 90% (traitillé), et distribution "mismatch" moyenne obtenue dans le cas d'une seule expansion (en gris). ***S1, C1*** et ***C2*** = principales périodes de coalescence, voir texte.

4.5.3.1 Influence de $N_{AG}m$:

Tout comme dans le cas d'une unique expansion spatiale (section 3.2), un petit $N_{AG}m$ (<50) provoque des coalescences récentes durant la "scattering phase" (Wakeley 1999) – pour laquelle nous utiliserons dorénavant l'abbréviation *S1* – qui n'existent quasiment pas pour un grand $N_{AG}m$ (Figure 4.14A). Ces coalescences récentes se traduisent par une classe 0 importante dans les distributions "mismatch" (Figure 4.14C). La plupart des coalescences qui n'ont pas lieu pendant la période *S1* lorsque $N_{AG}m$ est important ont lieu pendant la période *C1* ("contraction 1"). *C1* n'existe pas lors d'une expansion simple et correspond aux coalescences provoquées par le goulet néolithique. Les dernières coalescences se font pendant la période *C2* ("contraction 2"), qui correspond à la phase de contraction de la vague paléolithique.

En comparant les distributions "mismatch" obtenues avec deux expansions et celles obtenues avec une seule expansion (en gris dans la Figure 4.14C.), on peut conclure que l'homozygotie attendue est systématiquement plus importante dans le premier cas. En effet, même avec un grand $N_{AG}m$ (1'000), un premier mode est observé lorsque deux expansions se succèdent avec un remplacement complet des *CC*, ce qui n'est pas le cas avec une expansion unique. Il s'agit d'une différence importante entre les expansions spatiales et les simples croissances démographiques. Il est en effet impossible de distinguer la signature génétique obtenue après deux croissances démographiques qui se succèdent dans une population non-subdivisée, de celle obtenue après une seule croissance. Il est nécessaire que $N_{AG}m$ soit plus petit que 2 pour que la trace de l'expansion paléolithique disparaisse complètement (résultat non montré). Dans tous les autres cas, la signature génétique de l'expansion paléolithique est observable dans la distribution "mismatch" moyenne, sous la forme du mode le plus à droite. Cependant, cette signature n'est pas forcement décelable lors d'une observation unique car la variance est grande, particulièrement lorsque $N_{AG}m$ est petit.

4.5.3.2 Influence de $N_{CC}m$

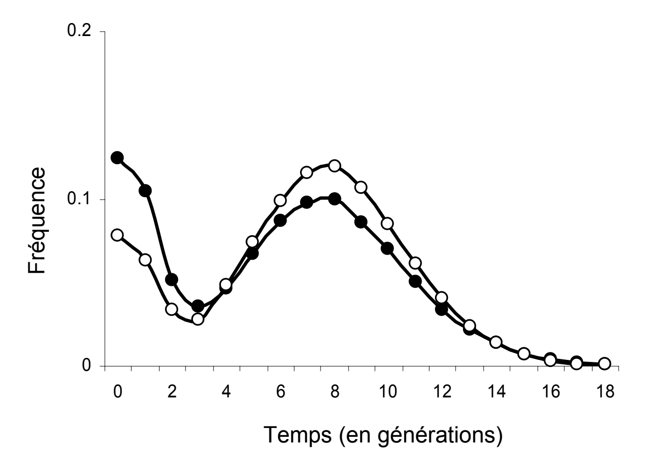


Figure 4.15 Distributions "mismatch" moyennes obtenues pour un petit (= 10, cercles noirs) et un grand $N_{CC}m$ (= 100, cercles blancs). Avec un $N_{AG}m$ = 1'000, identique dans les deux cas.

Plus $N_{CC}m$ est petit et plus l'homozygotie attendue est élevée (Figure 4.15), même si cette augmentation est beaucoup plus faible que celle provoquée par $N_{AG}m$. Ceci est dû à l'augmentation des coalescences provoquée par la concentration des lignages lors de leur passage par le goulet néolithique, comme le montre la Figure 4.16. Cette figure montre que lorsque $N_{CC}m$ est petit (= 10), environ 30% des gènes qui passent le goulet coalescent pendant les 50 générations qui précèdent celui-ci, alors que lorsque $N_{CC}m$ est grand (=100), il n'y a pratiquement aucune coalescence pendant ce laps de temps (< 0.4%). Plus $N_{CC}m$ est petit et plus la probabilité de coalescence est donc importante par rapport à la probabilité de dispersion, à la suite du goulet. L'influence de $N_{CC}m$ est tout de même nettement moindre que celle de $N_{AG}m$.

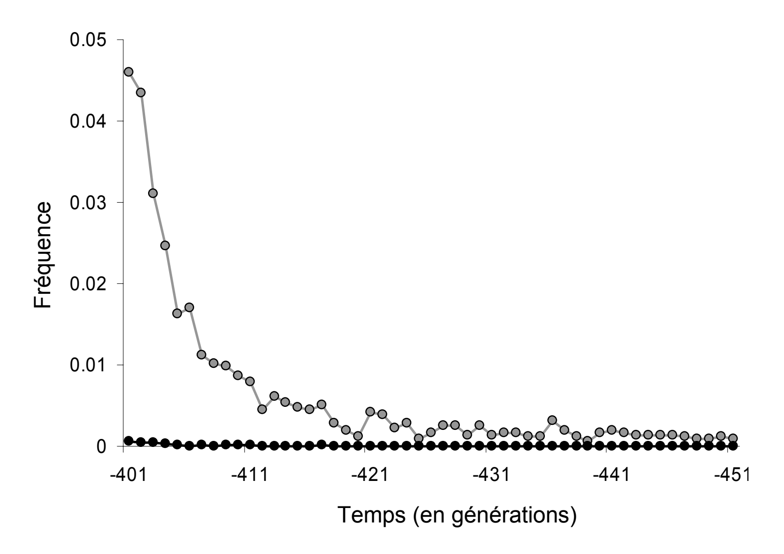


Figure 4.16 Distribution des coalescences pendant les générations qui précèdent le goulet néolithique (t = -400), pour $N_{CC}m$ = 10 (gris) et 100 (noirs), avec $N_{AG}m$ = 1'000 identique dans les deux cas.

4.5.3.3 Influence des taux de croissance r_{AG} et r_{CC}

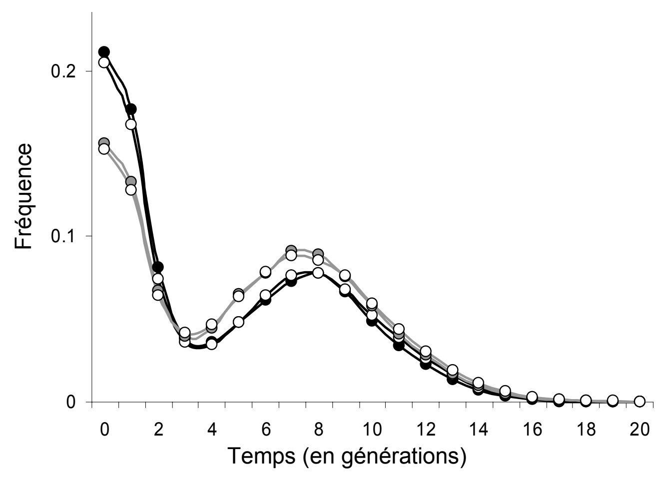


Figure 4.17 Distributions "mismatch" obtenues avec $N_{CC}m$ = 10 et $N_{AG}m$ = 1'000 et différents taux de croissances : r_{AG} = 0.1 (trait noir) et = 0.5 (trait gris). r_{CC} = 0.1 (cercles pleins) et 0.5 (cercles vides).

Lorsque le taux de croissance néolithique r_{AG} est grand, alors la colonisation du monde se fait plus rapidement (Ray 2003: p. 190), ce qui implique une réduction de l'homozygotie attendue,

puisque la probabilité de coalescence des lignages est plus faible. Cet effet est cependant nettement moindre que celui de $N_{AG}m$, comme le montre la Figure 4.17. Il avait déjà été constaté qu'une ré-expansion rapide après un goulet d'étranglement diminue l'effet de ce dernier, puisque les allèles n'ont pas le temps de disparaître par dérive (Chakraborty 1975). L'effet de r_{CC} est, quant à lui, négligeable (Figure 4.17).

4.5.3.4 Influence du goulet d'étranglement

Nous avons également étudié l'effet de la taille du goulet d'étranglement ("bottleneck" en anglais) sur les généalogies. Les deux aspects du goulet que nous étudions ici sont : *i)* sa taille, soit le nombre d'individus qui le constitue, *ii)* sa position par rapport au lieu d'échantillonnage. Dans les simulations qui suivent, nous faisons l'hypothèse que $N_{AG}m$ est grand (= 1'000) et que $N_{CC}m$ est petit (= 10). Les autres paramètres utilisés sont r_{CC} = 0.3, r_{AG} = 0.5 et m = 0.1.

- Taille du goulet :

Plus le nombre d'individus qui passent par le goulet d'étranglement est grand et moins l'homozygotie attendue est importante (Figure 4.18), puisque les coalescences de la période *C1* sont moins nombreuses. Cette réduction est très importante lorsque l'échantillonnage est fait à l'endroit même du lieu d'origine de l'expansion néolithique (*L* pour local), mais elle est nettement moindre lorsque l'échantillonnage est fait en périphérie de celle-ci (*P*). En effet, lorsque l'échantillonnage a lieu en périphérie, de nombreuses coalescences se passent avant d'arriver au goulet d'étranglement (voir ci-dessous) et la taille de ce dernier a donc une importance moindre.

- Localisation du goulet :

La localisation d'un échantillon par rapport au lieu de l'expansion de la population, peut avoir une certaine influence sur la signature génétique. En effet, des barrières géographiques telles que des montagnes ou des côtes peuvent avoir un effet sur la dispersion des gènes et ainsi modifier leurs migrations. Il nous a donc paru important de voir, dans un premier temps, si les effets de bord induits par le lieu d'échantillonnage peuvent jouer un rôle dans la signature génétique d'une double expansion. Nous avons donc déplacé l'expansion néolithique en périphérie du monde virtuel, dans le dème <5 ; 5>, afin d'avoir un lieu d'échantillonnage (dème <25 ;25>) qui ne soit pas localisé au même endroit. Comme le montre la Figure 4.19, la dynamique spatiale des gènes est passablement affectée lorsque le lieu d'échantillonnage est différent du lieu d'origine de l'expansion néolithique (en comparaison d'un échantillonnage local, Figure 4.13). En effet, les gènes qui sont ramenés vers le goulet sont rassemblés dans des dèmes communs au front de la vague de migration, et par conséquent le nombre de coalescences augmente avant d'arriver au goulet. L'homozygotie attendue d'une population est donc plus importante en périphérie qu'à l'endroit même du goulet (Figure 4.18). Cette différence suggère qu'il doit être possible de localiser la position d'une expansion en analysant séparément la diversité intrapopulationnelle d'échantillons indépendants et localisés à des endroits différents. Il faudrait néanmoins des distributions "mismatch" obtenues à partir de nombreux locus indépendants pour pouvoir en tirer quelques informations ; la comparaison

de deux distributions "mismatch" obtenues pour des échantillons différents, à partir d'un seul locus, ne peut donner aucune indication fiable du fait de sa très grande variabilité.

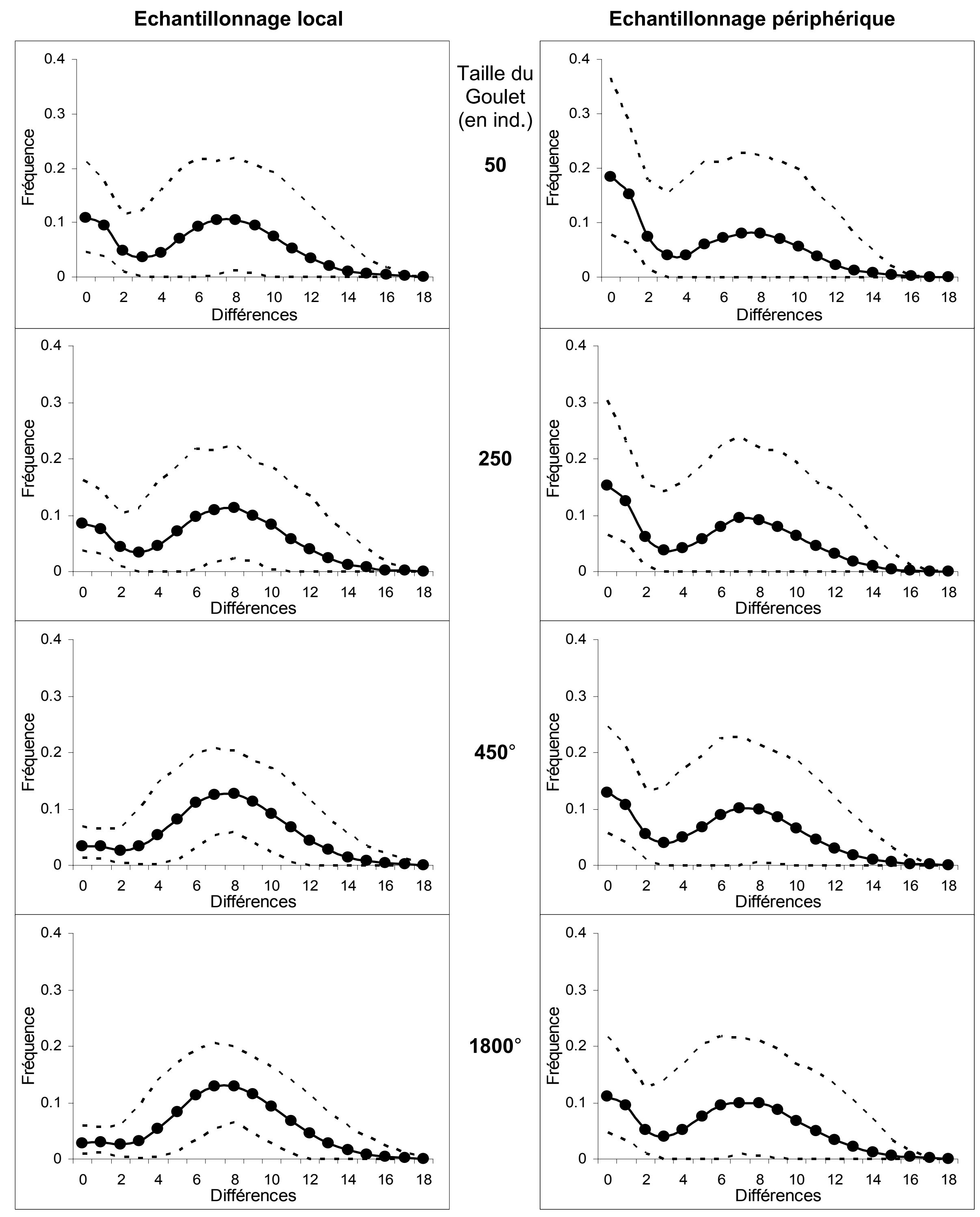


Figure 4.18 Distributions "Mismatch" moyennes pour 1'000 simulations et intervalle de confiance à 90% (en traitillés) pour différentes tailles du goulet (en individus). Colonne de gauche : échantillonnage sur le lieu du goulet; Colonne de droite : échantillonnage en périphérie. $N_{CC}m$ = 10, $N_{AG}m$ = 1'000, r_{CC} = 0.3, r_{AG} = 0.5 et m = 0.1. ° goulet constitué de 9 dèmes à la place d'un seul.

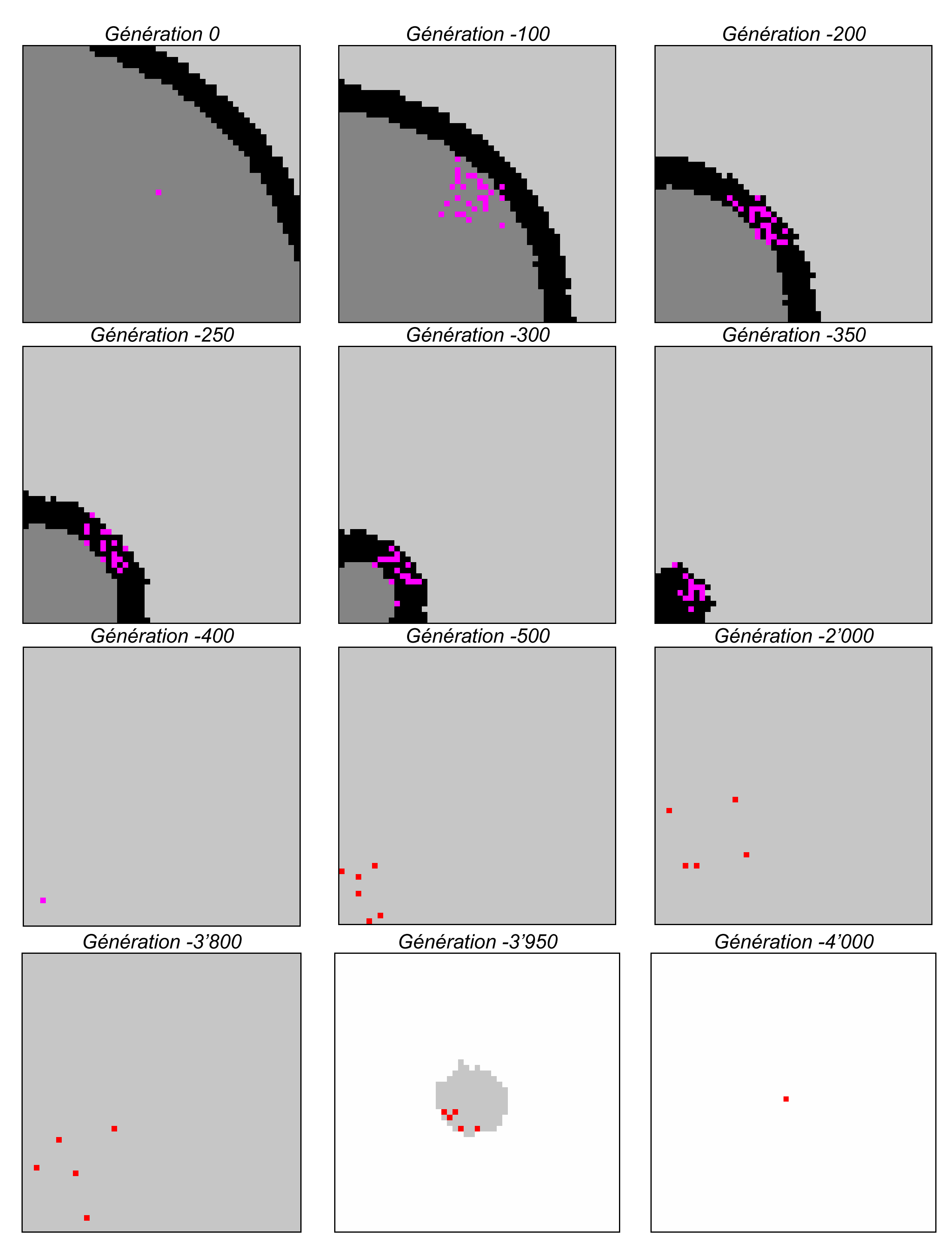


Figure 4.19 Occupation du monde : gris claire = occupation par *CC*, gris foncé = occupation par *AG*, noir = occupation par *CC* et *AG*. Rose = dème de la population *AG* dans lequel se trouve au moins un gène; Rouge = dème de la population *CC* dans lequel se trouve au moins un gène.

4.5.3.5 Influence du taux d'hybridation γ

Lorsque des migrations interpopulationnelles (hybridation) sont simulées depuis les chasseurs-collecteurs vers les agriculteurs (*CC* → *AG*), la proportion de gènes échantillonnés dont les ancêtres sont issus du dème source de la population *AG* (les "gènes néolithiques") diminue de façon exponentielle avec l'augmentation de γ (Figure 4.20). Cette diminution est d'autant plus importante que l'échantillonnage est fait en périphérie de la source de la population néolithique. Localement, il y a toujours au moins 5% des gènes qui sont issus de la population néolithique originale, même lorsque le taux d'hybridation est à son maximum (γ = 1). Ces observations suggèrent que si le patrimoine génétique européen actuel est composé d'une large fraction de gènes issus des premiers agriculteurs du Proche-Orient, alors la contribution indigène lors du Néolithique ne peut avoir été que très faible, voire nulle.

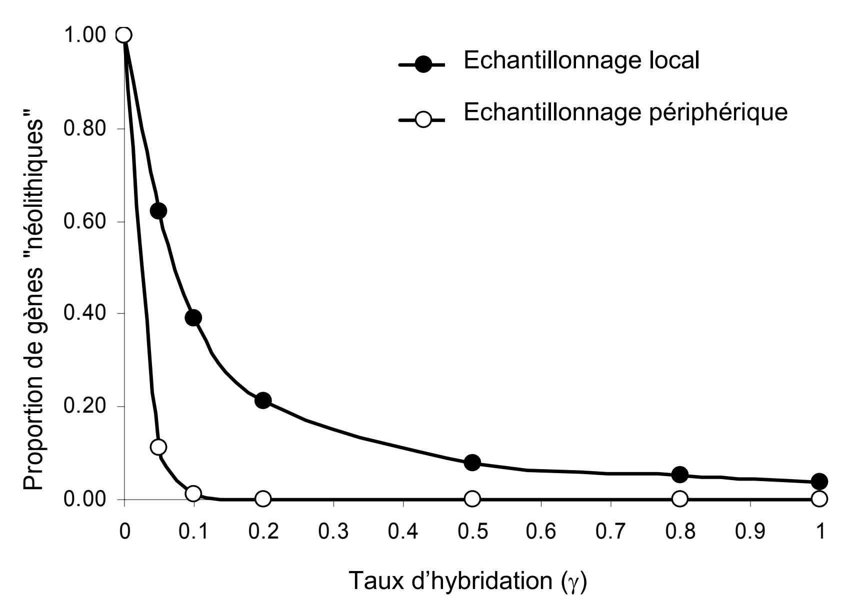


Figure 4.20 Proportion de "gènes néolithiques", dont les ancêtres sont issus de la population *AG* initiale, en fonction de γ, pour un échantillonnage sur le lieu de l'expansion néolithique (en noir) ou en périphérie (en blanc).

Nous avons montré que l'une des conséquences d'un remplacement des chasseurs-collecteurs lors de la transition néolithique est une homozygotie attendue importante dans les populations, due aux effets combinés des coalescences des périodes *S1* et *C1* (Figure 4.14). Cette homozygotie attendue, visible dans les distributions "mismatch", disparaît avec l'augmentation de γ d'autant plus vite que l'échantillonnage est effectué en périphérie de la source néolithique. En effet, c'est dans le front d'avancée néolithique que se passe la cohabitation entre agriculteurs et chasseurs-collecteurs et c'est également pendant cette cohabitation que de l'hybridation est possible. Par conséquent, plus les gènes se trouvent longtemps dans le front de la vague d'avancée et plus leur probabilité d'être issus de la population *CC* est grande. Lorsque γ est suffisamment grand (> 0.1), alors les distributions "mismatch" obtenues lors d'une double expansion sont identiques à celles obtenues avec une simple expansion (résultats non montrés). Cela implique qu'il suffit d'une faible

incorporation indigène lors de chaque étape de la progression des techniques agropastorales, pour qu'aucune trace spécifique au Néolithique ne soit visible dans la diversité intradème des populations européennes.

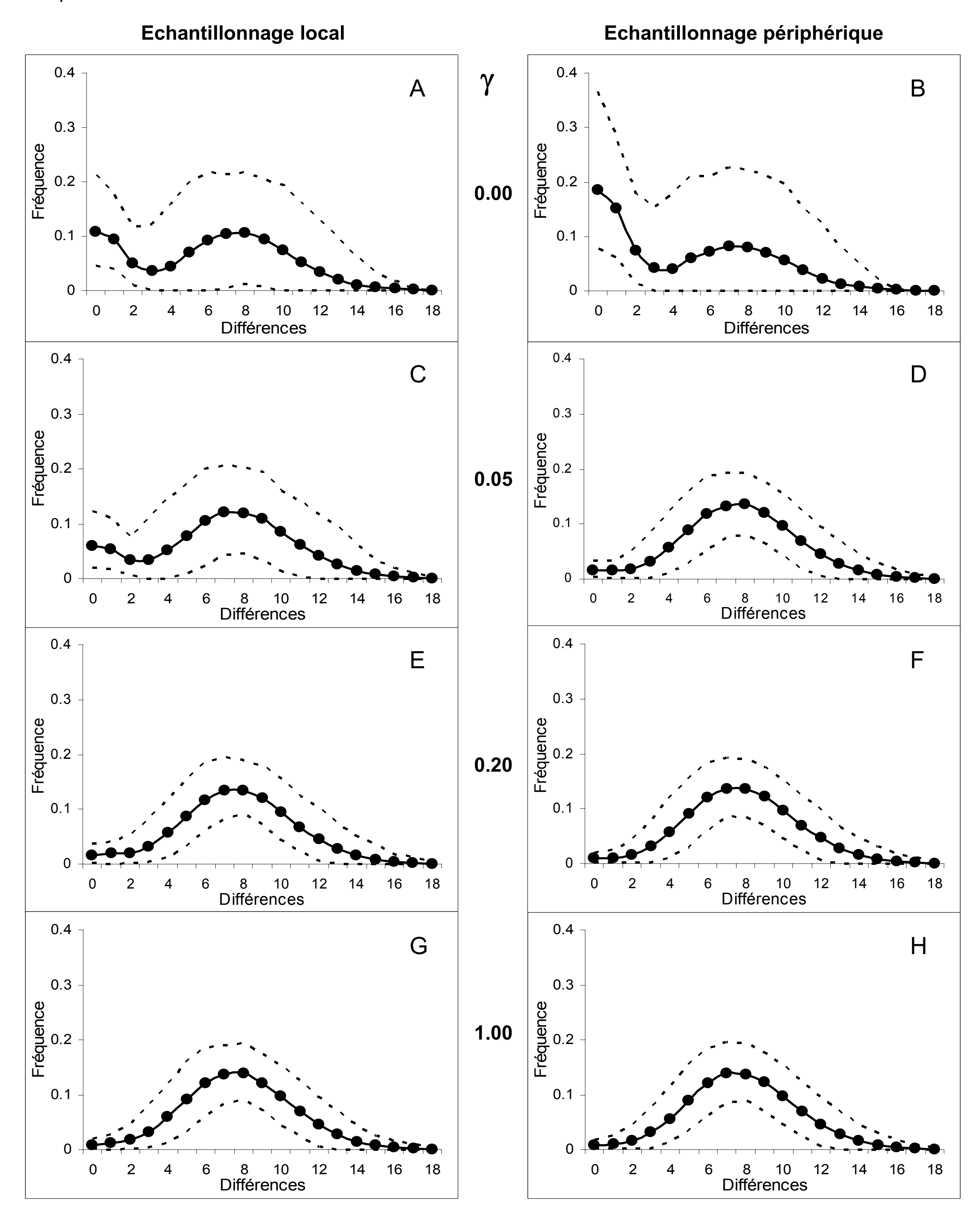


Figure 4.21 Distributions mismatch moyennes simulées dans le cas d'un échantillonnage sur le lieu d'oirigne de l'expansion néolithique ou en périphérie de celle-ci, lorsque le taux d'hybridation γ varie. En traitillé l'intervalle à 90%. $N_{CC}m$ = 10, $N_{AA}m$ = 1'000, r_{CC} = 0.3, r_{AA} = 0.5 et m = 0.1.

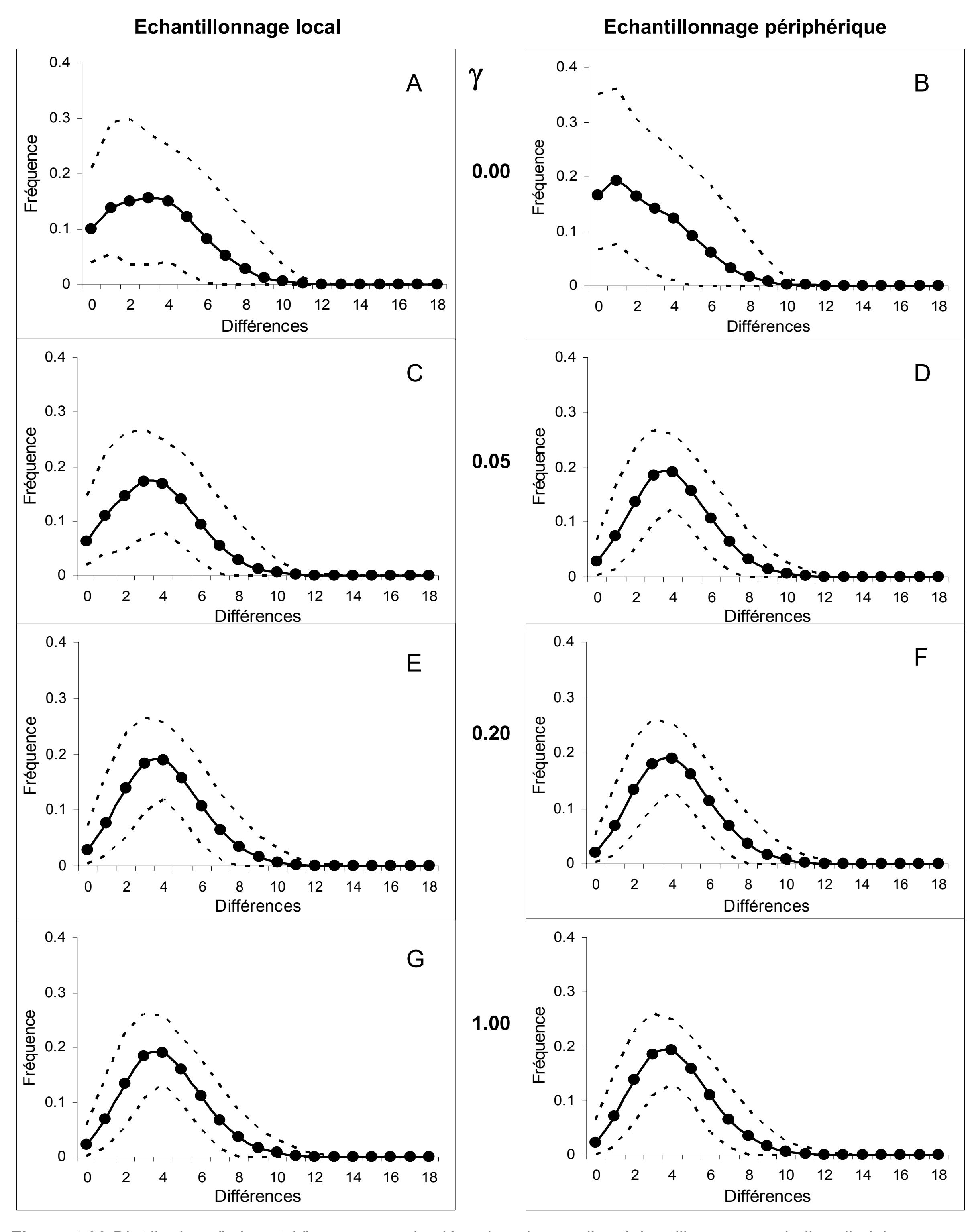


Figure 4.22 Distributions "mismatch" moyennes simulées dans le cas d'un échantillonnage sur le lieu d'origine de l'expansion néolithique ou en périphérie de celle-ci, lorsque le taux d'hybridation γ varie. En traitillé l'intervalle à 90%. $N_{CC}m$ = 10, $N_{AG}m$ = 1'000, r_{CC} = 0.3, r_{AG} = 0.5 et m = 0.1.

4.5.3.6 Cadre temporel et taux de mutation

Jusqu'ici nous avons simulé une première expansion correspondant à la diffusion initiale d'*Homo sapiens sapiens*, datée d'environ 100'000 ans (4'000 générations), notamment pour des raisons de clarté de l'exposé. Le peuplement du continent européen est plus récent et date d'environ 40'000 ans, peut-être depuis le Proche-orient (Stringer 1989) ou depuis une autre source en Asie de l'ouest ou en Asie centrale (Djindjian *et al.* 1999 ; Kozlowski et Otte 2000). On peut donc légitimement se poser la question de l'influence que pourrait avoir sur les données moléculaires une expansion spatiale paléolithique plus récente que celle que nous avons simulée jusqu'ici. Nous avons donc procédé à une nouvelle série de simulations identique à la précédente, mais en datant la première expansion de -1'600 générations (à la place de 4'000), tout en gardant l'expansion néolithique à -400 générations. Nous avons également modifié légèrement le taux de mutation (μ = 0.00125 à la place de 0.001), afin que le mode attendu des distributions soit égal à 4, nombre qui correspond à l'ordre de grandeur de la majorité des distributions "mismatch" obtenues pour les populations européennes (Di Rienzo et Wilson 1991 ; Bertranpetit *et al.* 1995 ; Sajantila *et al.* 1995 ; Calafell *et al.* 1996 ; Comas *et al.* 1996 ; Corte-Real *et al.* 1996 ; Francalacci *et al.* 1996 ; Comas *et al.* 2000 ; Malyarchuk et Derenko 2001 ; Nasidze et Stoneking 2001).

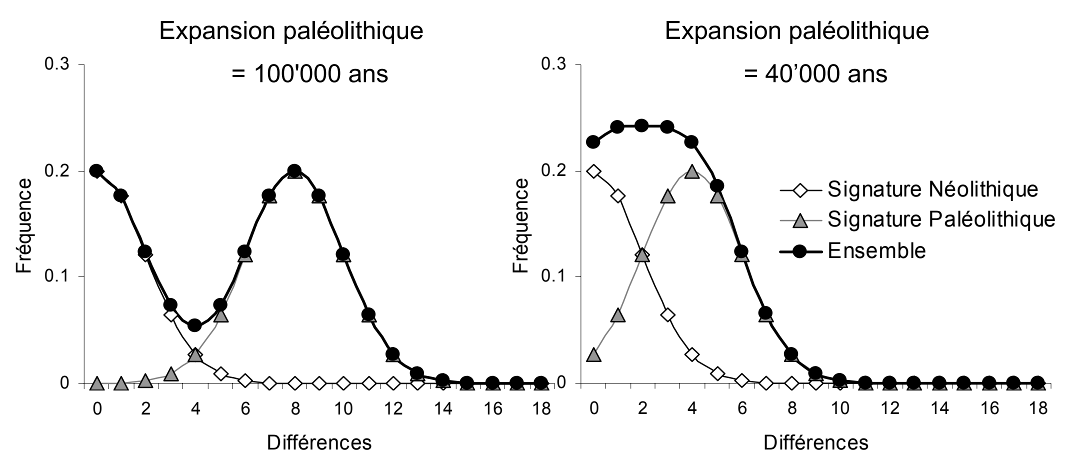


Figure 4.23 Distributions "mismatch" (cercle noir), lors d'un remplacement complet de la population *CC*, sous l'effet combiné des signatures d'expansions néolithique (losanges blancs) et paléolithique (triangle gris). A. = expansion paléolithique il y a 4'000 générations (~100'000 ans) ; B. = expansion paléolithique il y a 1'600 générations (~40'000 ans).

Toutes les observations faites précédemment et relatives aux effets des paramètres démographiques sur les généalogies sont toujours valables lorsque le cadre temporel est modifié (résultats non montrés). Une expansion plus récente de la population *CC* rapproche la période *S1* de la période *C2* (Figure 4.14). Il en résulte un raccourcissement des arbres de coalescence, dont la topologie générale reste semblable à celle observée lors d'une expansion paléolithique ancienne, puisque elle est indépendante du taux de mutation. En revanche, la forme des distributions "mismatch" est passablement différente lorsque l'hybridation est très faibles ou nulle (γ < 0.1). La Figure 4.22 présente les distributions "mismatch" obtenues avec les mêmes paramètres que celles illustrées par la Figure 4.21, mais cette fois dans le cadre d'une expansion récente de la population *CC* (-1'600 générations). Lorsque γ < 0.1, la distribution "mismatch" moyenne est unimodale et le

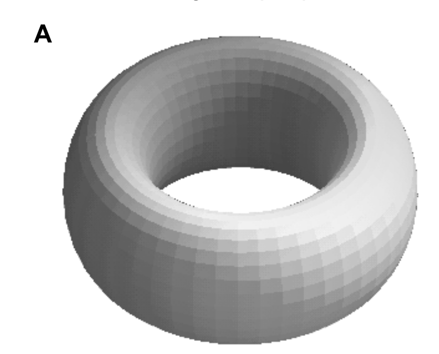
mode centré sur la classe 0 disparaît. Les valeurs des petites classes sont toutes relativement importantes et il n'est plus possible d'observer de distribution "mismatch" moyenne bimodale. Lorsque l'hybridation est importante (γ > 0.1), alors des distributions "mismatch" unimodales sont observées dans tous les cas, quelque soit le taux de mutation (μ) ou le cadre temporel (τ_C) utilisé.

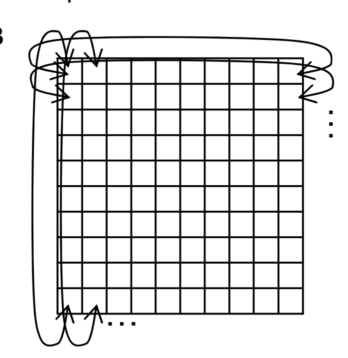
Dans le cas d'un remplacement important des chasseurs-collecteurs, chacun des deux modes est la signature d'une expansion démographique différente : le premier mode (à gauche) est la signature de l'expansion néolithique et le second, celle de l'expansion paléolithique (Figure 4.23A). Lorsque ces deux expansions sont très proches temporellement, comme c'est le cas avec une expansion paléolithique il y a 1'600 générations, il n'est plus possible de discerner leur signature dans les distributions "mismatch" moyennes (voir Figure 4.23B).

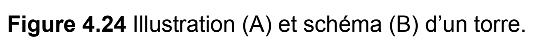
Il faut donc faire très attention à l'interprétation de la forme des distributions "mismatch" dans les populations réelles. Non seulement leur variance est grande, particulièrement lorsque γ est faible, mais de plus leur forme peut être modifiée par le cadre temporel ou par le taux de mutation.

4.5.3.7 Forme du monde

Afin d'éviter que les résultats présentés ne soient le fruit d'effets de bord dus à la forme de notre monde, nous avons doublé toutes les simulations effectuées jusqu'ici en utilisant un monde, non plus carré, mais en forme de torre (Figure 4.24). Les cellules situées de chaque côté du monde peuvent communiquer avec les cellules du côté opposé. Les résultats observés montrent des différences que l'on peut considérer comme négligeables lorsque γ > 0. En revanche lorsque γ = 0 et que le lieu d'échantillonnage est différent de l'origine de l'expansion, on observe une légère réduction de l'homozygotie attendue. Les valeurs observées sont intermédiaires entre celles obtenues, dans le monde carré, pour un lieu d'échantillonnage différent du lieu d'expansion et celles obtenues pour un lieu d'échantillonnage identique au lieu d'expansion. Ceci s'explique facilement par le fait que dans un monde en torre, les gènes échantillonnés ont plus de possibilités de migration que dans un monde carré, ils auront donc moins tendance à se retrouver dans les mêmes dèmes au front de la vague d'expansion, et les coalescences qui ont lieu entre la période *S1* et *C1* lors d'un échantillonnage en périphérie du lieu du goulet sont moins importantes.






Figure 4.24 Illustration (A) et schéma (B) d'un torre.

4.5.4 Discussion

Nous avons montré dans la section précédente (4.5.3), qu'il existe trois périodes principales où se produisent des coalescences (*S1*, *C1* et *C2*, Figure 4.25) lorsque deux expansions démographiques et spatiales se succèdent dans la même aire. Le goulet d'étranglement qui est créé au début de la seconde expansion (néolithique) donne naissance à une troisième période (*C1*) propice aux coalescences, qui n'existe pas dans le cas d'une expansion unique (Figure 4.25). Le nombre de coalescences qui a lieu durant cette période *C1* est plus ou moins grand selon l'importance des paramètres démographiques, dont les quatre plus influents sont, par ordre d'importance, la valeur du paramètre *Nm* de la seconde population ($N_{AG}m$), l'hybridation, la localisation de l'échantillon par rapport au goulet, et le nombre d'individus qui compose ce dernier. Les autres paramètres n'ont qu'une influence plus modeste.

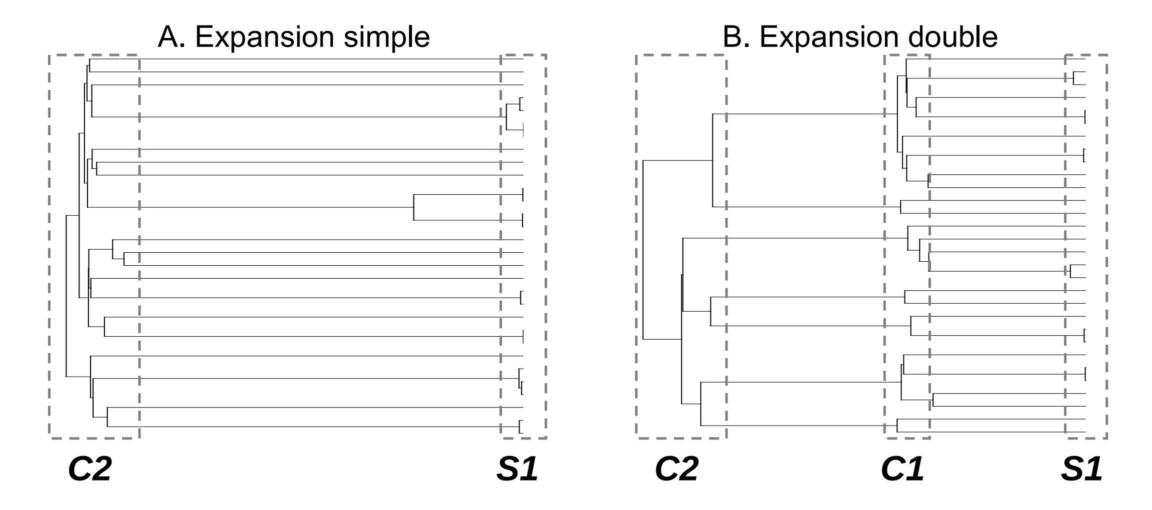


Figure 4.25 Exemples de généalogies de gènes obtenus dans le cas d'une simple (A) ou d'une double (B) expansion démographique. *S1* = coalescences récentes ("*scattering phase*"); *C1* = coalescences au moment de l'expansion néolithique (*"contraction 1*") ; *C2* = coalescences au moment de l'expansion paléolithique (*"contraction 2*").

Le temps très court (400 générations) qui s'écoule entre l'expansion néolithique et le présent conduit à des branches terminales de l'arbre qui sont aussi très courtes. Les mutations n'ont pas le temps de s'accumuler sur ces branches, pas plus qu'elles n'ont le temps de s'accumuler sur les branches terminales qui se font à partir des coalescences de la période *S1*. Ceci implique que les signatures génétiques intrapopulationnelles de ces deux périodes de coalescences s'additionnent et ne permettent pas de les différencier. La signature des périodes de coalescence *S1* et *C1* sur la distribution "mismatch" est l'apparition d'une importante classe 0 (homozygotie), qui correspond au premier mode de la distribution. Les distributions "mismatch" obtenues lors d'un remplacement complet de la première population (paléolithique) sont donc identiques à celles obtenues par une unique expansion dans une population de faible *Nm* (section 3.2). En revanche, lorsque la contribution génétique de la population paléolithique est importante, les distributions "mismatch" simulées sont semblables à celles obtenues après la diffusion d'une seule population de grand *Nm*. Or, il suffit d'une faible contribution indigène dans chaque dème pour que leur contribution globale

soit très importante. Par conséquent, pour que la signature génétique de l'expansion néolithique soit décelable dans la population actuelle, il faut obligatoirement que le remplacement de la population indigène ait été très fort. A l'inverse, la signature génétique de l'expansion paléolithique dans la population actuelle ne disparaît jamais, quelle que soit l'importance du goulet néolithique et de la contribution indigène.

Lorsque plusieurs croissances démographiques simples se succèdent dans une population non-subdivisée, il est impossible de différencier leur signature génétique respective. A l'inverse, nous avons montré, à l'aide de nos simulations, que deux expansions démographiques et spatiales qui se succèdent peuvent être distinguées dans la structure génétique des populations. Cependant, plus ces expansions spatiales sont proches temporellement et plus il est difficile de différencier leurs signatures génétiques, puisque le résultat des coalescences qui ont lieu aux différentes périodes (*S1*, *C1* ou *C2*) se confondent. Les modes générés dans la distribution "mismatch" par chacune des deux expansions fusionnent (Figure 4.23). Plus deux expansions sont proches temporellement, et plus il est nécessaire d'utiliser des locus ayant un fort taux de mutation pour pouvoir distinguer leur trace dans les données moléculaires intrapopulationnelles.

Ces résultats suggèrent que l'interprétation des distributions "mismatch" des populations humaines doit être faite avec une grande prudence puisque d'une part, la variance de ces distributions peut-être grande – particulièrement dans les populations de faible densité – et d'autre part différents scénarios démographiques peuvent donner des distributions "mismatch", ainsi que des statistiques intradèmes, très semblables. Malgré cela, nous avons montré dans ce chapitre, que les distributions "mismatch" unimodales observées dans la majorité des populations post-néolithiques européennes (Figure 4.26) ne peuvent pas résulter d'un remplacement complet des chasseurs-collecteurs à la période néolithique. Nos simulations ont en effet montré que lors d'un remplacement complet des chasseurs-collecteurs, un premier mode centré sur la classe 0 est toujours observé dans les populations périphériques à la zone d'origine du Néolithique. Cette observation est indépendante de la taille initiale de la population néolithique (Figure 4.18). De plus, la régularité avec laquelle ces distributions "en cloche" sont observées dans les populations européennes (Di Rienzo et Wilson 1991 ; Bertranpetit *et al.* 1995 ; Sajantila *et al.* 1995 ; Calafell *et al.* 1996 ; Comas *et al.* 1996 ; Corte-Real *et al.* 1996 ; Francalacci *et al.* 1996 ; Comas *et al.* 2000 ; Malyarchuk et Derenko 2001 ; Nasidze et Stoneking 2001) suggère même une contribution génétique indigène importante lors du Néolithique, puisque la variance des distributions "mismatch" diminue avec l'incorporation de chasseurs-collecteurs dans la population néolithique. Il n'est cependant pas possible d'estimer avec précision la contribution des chasseurs-collecteurs au patrimoine génétique européen à l'aide des distributions "mismatch" tirées du génome mitochondrial des populations actuelles. Il est en effet très difficile de comparer quantitativement des distributions "mismatch", d'autant plus que leur variance est grande et qu'elles ne sont tirées que d'un seul locus.

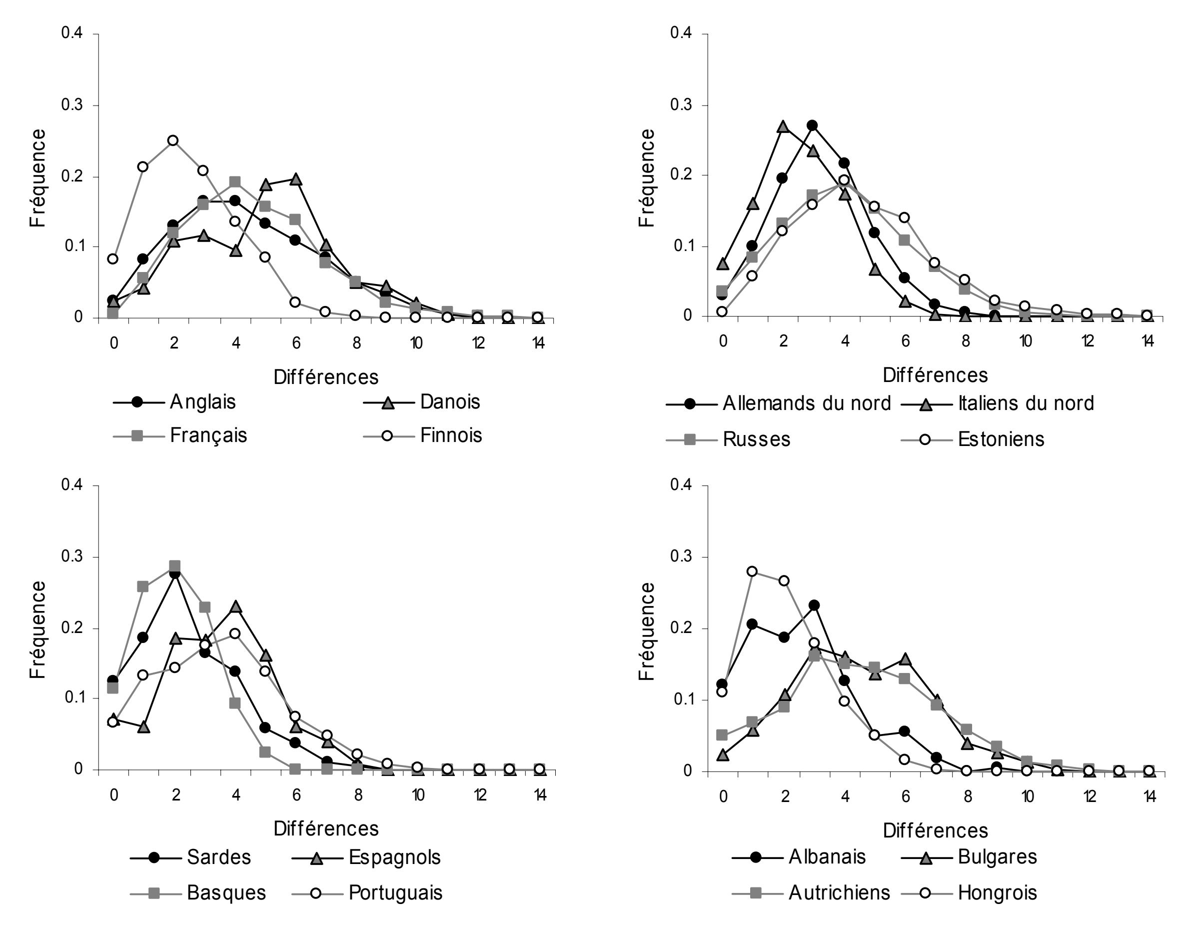


Figure 4.26 Exemples de distributions "mismatch" observées dans certaines populations européennes. *n* = 100 Anglais (Piercy *et al.* 1993), 33 Danois, 106 Allemands du nord (Richards *et al.* 1996), 50 Français (Rousselet et Mangin 1998), 50 Finnois, 28 Estoniens (Sajantila *et al.* 1995), 68 Italiens du nord (Mogentale-Profizi *et al.* 2001), 103 Russes (Orekhov *et al.* 1999), 69 Sardes (Di Rienzo et Wilson 1991), 41 Espagnols, 54 Portuguais (Corte-Real *et al.* 1996), 45 Basques (Bertranpetit *et al.* 1995), 42 Albanais (Belledi *et al.* 2000), 30 Bulgares (Calafell *et al.* 1996), 101 Autrichiens (Parson *et al.* 1998), 35 Hongrois (Kalmar *et al.* 2003).

4.6 Conclusion

Dans ce chapitre, nous avons présenté le développement d'une méthodologie qui permet de simuler les interactions entre deux populations évoluant dans une aire géographique donnée. Deux types d'interactions sont possibles entre ces populations : des échanges génétiques et de la compétition. Si des modèles démographiques avaient déjà été développés pour simuler un flux génétique entre deux populations (Rendine *et al.* 1986 ; Barbujani *et al.* 1995 ; Aoki 1996 ; Aoki *et al.* 1996) ou de la compétition entre elles (Flores 1998), notre approche est la première qui prenne en compte les deux types d'interactions simultanément.

L'implémentation de notre modèle dans une version modifiée du logiciel SPLATCHE (voir ANNEXE 4) offre de nombreux avantages, parmi lesquels la possibilité de tester de façon intensive les paramètres du modèle, de visualiser les composantes démographiques et génétiques, et surtout d'utiliser ce modèle dans une infinité de cadres temporels et géographiques différents.

Nous avons développé cette méthodologie dans le but de simuler l'expansion d'une population dans une aire préalablement peuplée et plus particulièrement le remplacement des Néandertaliens par *Homo sapiens sapiens* (chapitre 5), ainsi que la transition néolithique européenne (chapitre 6). Avant de procéder à ces recherches, il était important de bien cerner le comportement du modèle et l'influence des différents paramètres impliqués sur les données génétiques. Cette exploration du modèle a été effectuée dans le contexte des populations humaines selon un schéma identique à celui utilisé dans le chapitre 3. De cette manière, nous avons pu cerner les différences qui existent dans la diversité moléculaire d'une population qui a colonisé une aire vide ou une aire déjà peuplée. Nous avons ainsi montré qu'il suffit d'une très faible incorporation d'individus appartenant à la population indigène dans chaque dème, pour que la signature moléculaire obtenue lors de la colonisation d'une aire peuplée ressemble à celle obtenue lors de la colonisation d'une aire vide. En revanche, si le remplacement de la population indigène est complet, ou presque complet, alors les signatures résultant de chacune des deux expansions spatiales sont visibles dans la structure génétique. Ces signatures se traduisent par deux modes distincts dans les distributions "mismatch", qui résultent des coalescences très fréquentes qui ont lieu pendant les périodes d'expansion. Il faut préciser que plus les deux expansions sont proches temporellement et plus il devient difficile de distinguer leurs signatures respectives.

Appliquées aux populations européennes, ces observations suggèrent que la transition néolithique sur ce continent n'a pas pu se faire avec un remplacement très important de la population indigène, comme proposé par le modèle initial de diffusion démique (Ammerman et Cavalli-Sforza 1984). En effet, si le remplacement des chasseurs-collecteurs mésolithiques avait été important pendant le Néolithique, on s'attendrait à trouver une classe 0 importante, ainsi qu'une grande variance, dans les distributions "mismatch" tirées des populations européennes, ce qui n'est pas le cas. Cette comparaison ne porte cependant que sur l'inspection visuelle des données provenant d'un seul locus et ne permet donc pas une estimation précise de la contribution indigène lors du Néolithique.

Notre méthodologie peut être également utilisée dans de nombreux autres contextes que ceux qui sont présentés dans ce chapitre ou dans les chapitres 5 et 6, notamment dans le cadre d'études sur l'"incorporation" ("introgression") du génome d'une espèce envahie dans celui d'une espèce invasive (Bernatchez *et al.* 1995 ; Shaw 2002), sur la diffusion de nouvelles mutations (Klopfstein in prep.), ou sur d'autres espèces (ou sous-espèces) en compétition.

5 Expansion des Hommes modernes en Europe

5.1 Introduction

L'Europe et le Proche-Orient font partie des régions pour lesquelles l'histoire des Hommes est la plus abondamment documentée (voir par ex. :Ammerman et Cavalli-Sforza 1984 ; Renfrew 2000 ; Barbujani et Bertorelle 2001). Il est d'ailleurs impossible d'étudier le peuplement du continent européen sans prendre en considération les régions asiatiques qui le bordent, notamment le Proche-Orient, tellement leurs histoires sont intimement liées. C'est en effet depuis le sud-ouest de l'Asie que les principaux flux migratoires humains se sont faits en direction du continent européen. Les innombrables fouilles archéologiques menées soit en Europe, soit au Proche-Orient, ont permis la reconstitution plus ou moins précise de l'histoire de l'humanité dans ces régions (par ex. :Renfrew 1989 ; Mellars 1992 ; Gallay 1994 ; Whittle 1996 ; Djindjian *et al.* 1999 ; Mazurié de Keroualin 2003). Malgré cette abondance d'informations, il demeure encore énormément d'inconnues au sujet de l'histoire de nos ancêtres sur le continent européen. En effet, les restes archéologiques sont peu nombreux pour certaines périodes ou plus difficiles à interpréter, et il existe une grande hétérogénéité dans la couverture du continent (Hazelwood et Steele 2003). De plus, les techniques de datation ne permettent pas toujours des estimations précises et l'intervalle de confiance autour de ces dates peut être important (Gkiasta *et al.* 2003). Les paléontologues et les archéologues tentent néanmoins de reconstituer l'histoire de nos ancêtres en proposant des hypothèses de peuplement, dont certaines sont fortement débattues. Dans ce climat d'incertitude, la génétique apporte de nouveaux éléments de réponse à certaines interrogations soulevées par les autres disciplines. Le séquençage de portions d'ADN tirées d'os de Néandertaliens a notamment permis d'aborder sous un nouvel angle la question d'éventuels échanges génétiques entre ces derniers et les premiers Hommes modernes. Dans ce chapitre, nous essayons d'estimer le taux d'hybridation entre Néandertaliens et Hommes modernes qui est compatible avec les données génétiques actuelles. Nous utilisons pour cela l'approche par simulation présentée dans le chapitre précédent (4). Cette étude est présentée sous la forme d'un manuscrit soumis à publication.

5.2 Contribution des Néandertaliens au patrimoine génétique des Hommes modernes

Sur le continent européen, les successeurs d'*Homo erectus* ont développé des caractéristiques spécifiques qui culminent avec la forme classique d'*Homo neandertalensis*, dont l'apparition date d'environ -120'000 ans (Klein 2003). Les *H. neandertalensis* sont vraisemblablement endémiques à l'Europe occidentale et leur aire de répartition se serait petit à petit agrandie, jusqu'à atteindre le Proche-Orient il y a environ 80'000 ans (Figure 5.1, Hublin 1988 ; Klein 2003).

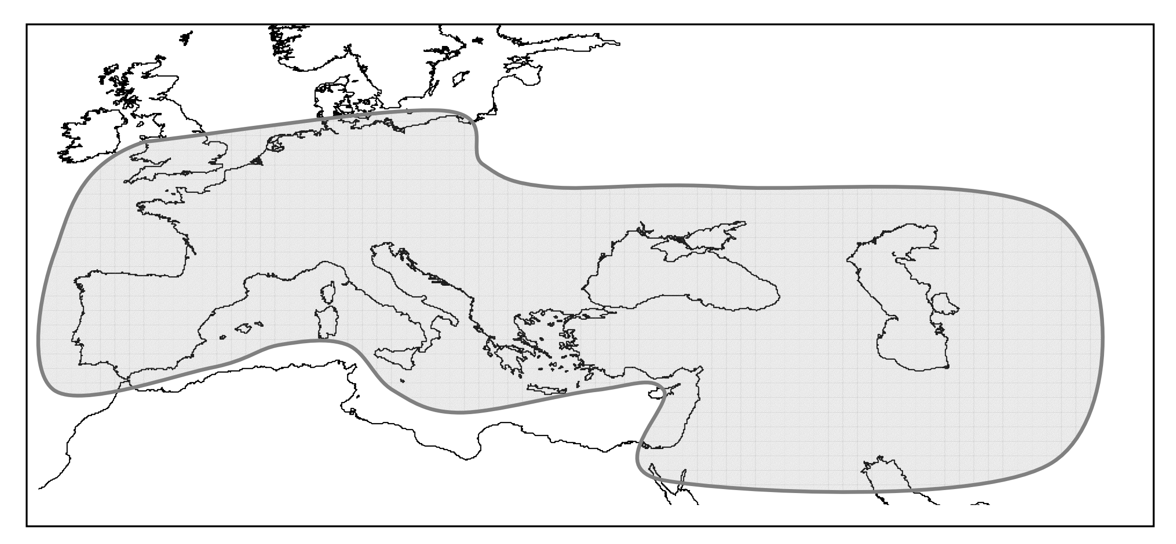


Figure 5.1 Aire de répartition approximative des Néandertaliens, figure modifiée à partir de Klein (2003).

C'est seulement entre 45'000 ans et 30'000 ans que les Hommes modernes (*Homo sapiens sapiens*) colonisent le continent européen (Mellars 1992 ; Mellars 1998 ; Otte 2000 ; Stringer et Davies 2001 ; Klein 2003) d'une part, à partir des plaines de Russie – qui ont servi de corridor – et d'autre part, à partir du Proche-Orient (Bocquet-Appel et Demars 2000a, 2000b). La présence de l'Homme moderne en Europe est attestée principalement par des vestiges lithiques appartenant à la culture aurigniacienne qui lui est associée. Cependant, "*l'origine exacte de l'Homme moderne européen n'est pas encore connue, le Moyen-Orient a été longtemps candidat mais l'Aurignacien du Levant y semble plus récent. L'Asie centrale est un candidat actuellement examiné mais il n'a pas été trouvé jusqu'à présent de site aurignacien antérieur à 35'000 BP*" (Djindjian *et al.* 1999: p. 162). La diffusion de l'Homme moderne en Europe a été extrêmement rapide et a sans doute contribué à l'extinction des *H.neandertalensis*, alors disséminés sur tout le continent. Bien que la coexistence entre les deux populations dure plusieurs milliers d'années (10'000 à 15'000 ans), il semble qu'il ne s'agisse pas d'une réelle cohabitation dans les mêmes lieux, mais plutôt d'une existence simultanée dans des régions voisines. Les *H.neandertalensis* ont, par exemple, occupé pendant près de 10'000 ans le sud de la péninsule ibérique, pendant que les *H.s.sapiens* peuplaient le nord de cette péninsule (Mellars 1998). Il faut préciser que la grande fluctuation de radiocarbone atmosphérique[1], entre 30'000 BP et 50'000 BP, rend difficile la mise en place de chronologies fiables pendant cette période, ainsi que l'estimation des temps de coexistence entre Hommes modernes et Néandertaliens (Conard et Bolus 2003). Les raisons qui ont mené à la disparition des Néandertaliens sont encore mal connues. Les *H.s.sapiens* ont sans doute joué un rôle dans cette extinction, peut-être en confinant les *H.neandertalensis* dans des environnements moins

---

[1] Le radiocarbone, aussi apellé carbone 14 ou C^{14}, est utilisé pour effectuer des datations allant jusqu'à environ -40'000 ans. L'âge de certains objets (stalagmite, corail, coquille, ossement, bois, charbon) peut être estimé en mesurant sa concentration en C^{14}, puisque celle-ci diminue de moitié tous les 5568 ans. Les dates ainsi obtenues (données en BP, pour Before Present) doivent cependant être ajustées, puisque la concentration de C^{14} dans l'atmosphère varie au cours du temps et que son accumulation dans les restes fluctue donc également. Ces dates "calibrées" sont données en BC (Before Christ), AD (Anno Domini) ou encore en "cal BP", et peuvent varier de plusieurs milliers d'années par rapport aux dates non calibrées (voir par exemple Grimaud-Hervé *et al.* 2001: pp. 114-117 ou le site "http://carbon14.univ-lyon1.fr/" pour plus de détails).

favorables : leur capacité d'adaptation étant réduite, ils n'auraient pas été capables de survivre aux variations climatiques extrêmement rapides à cette période (Bocquet-Appel et Demars 2000b ; Van Andel 2000 ; Stringer et Davies 2001 ; Hublin 2002). Si les échanges culturels semblent avoir été relativement restreints entre les deux groupes, excepté dans certaines régions (Klein 2003), la question de possibles échanges génétiques reste encore d'actualité (Hublin 1988 ; Duarte *et al.* 1999).

Le séquençage de fragments d'ADN ancien, tirés d'os de Néandertaliens, peut potentiellement apporter un élément de réponse à ce sujet. A ce jour, de courtes portions d'ADN mitochondrial de 8 individus néandertaliens ont pu être séquencées (Krings *et al.* 1997 ; Krings *et al.* 1999 ; Ovchinnikov *et al.* 2000 ; Schmitz *et al.* 2002 ; Serre *et al.* 2004). Leur comparaison avec les séquences d'Hommes modernes actuels (plus de 4'000 échantillons : Handt *et al.* 1998) a montré une très grande différenciation (Krings *et al.* 2000 ; Scholz *et al.* 2000), de même que leur comparaison avec des séquences d'*H.s.sapiens* de type Cro-Magnon (Caramelli *et al.* 2003 ; Serre *et al.* 2004). Cette grande différenciation a été interprétée comme la marque d'une hybridation faible, sinon nulle, entre les deux populations (Richards *et al.* 1996 ; Sykes 1999 ; Schillaci et Froehlich 2001). Cette dernière affirmation a cependant été contestée par Wall (2000) et Nordborg (1998), qui ont estimé qu'il est impossible de tirer une conclusion à propos du taux d'hybridation entre les deux espèces à partir d'un seul locus. Nordborg (1998) a d'ailleurs calculé que, même si les Néandertaliens composaient 25% du patrimoine génétique des premiers Hommes modernes européens, il existe 50% de chances pour que ces lignages aient été perdus par dérive génétique, en près de 40'000 ans (voir aussi Hagelberg 2003 et Relethford 2001). Récemment, avec un modèle plus réaliste que celui de Nordborg (1998) – incluant une croissance démographique des premiers Hommes modernes – Serre *et al.* (2004) ont estimé la contribution néandertaliène maximum à 25%, mais dont la valeur exacte est vraisemblablement faible et dépend des paramètres de l'expansion. Les calculs de Nordborg (1998) et de Serre *et al.* (2004) sont effectués sur la base d'un modèle de populations non-subdivisées et statiques d'un point de vue spatial.

Dans l'article présenté dans ce chapitre (section 5.2.1), nous avons simulé le remplacement des Néandertaliens par *Homo sapiens sapiens* à l'aide de la version modifiée de SPLATCHE et du modèle démographique présenté dans le chapitre 4. La diffusion des Hommes modernes est simulée pendant 1'600 générations (environ 40'000 ans) dans une matrice de dèmes homogènes représentant l'Europe, à partir d'un dème source situé au Proche-Orient. Une seconde matrice de dèmes, superposée à la première, est préalablement peuplée par des Néandertaliens en fonction de leur répartition (Figure 5.1). Ces derniers disparaissent sous l'effet de la compétition exercée par les Hommes modernes, dont la capacité de soutien doit être au moins 2,5 à 4 fois supérieure à celle des Néandertaliens pour que cette extinction ait lieu. Cette différence de densité est due à une meilleure exploitation des ressources par les Hommes modernes, probablement grâce à une technologie plus avancée ou à de meilleures facultés cognitives (Klein 2003). Nous avons simulé différentes proportions d'échanges génétiques entre les deux populations pendant une période de

cohabitation qui dure entre 7.5 et 12 générations (environ 200 à 300 ans) dans chaque dème de 2'500 km^2, en fonction des scénarios. En effet, 8 scénarios démographiques différents sont considérés, afin de tenir compte de l'incertitude liée au choix des valeurs de paramètres. Le scénario A est celui qui utilise les valeurs les plus probables estimées à partir des données de la littérature. Nous montrons dans cette étude que l'absence de lignage néandertalien dans la population européenne actuelle ne peut résulter que d'une hybridation extrêmement faible entre les deux populations concernées. Selon tous les scénarios simulés, la proportion initiale de gènes néandertaliens dans la population moderne n'a pu excéder 0.09% sans que l'on en observe encore des traces aujourd'hui. Cette estimation est environ 400 fois plus faible que celles faites préalablement (Nordborg 1998 ; Serre *et al.* 2004). Cette faible contribution peut être expliquée par le fait que l'hybridation se fait dans le front d'avancée de la vague de migration des Hommes modernes, au moment où ceux-ci sont encore peu nombreux. Ainsi, les rares hybrides néandertaliens vont contribuer à l'expansion des Hommes modernes et leurs gènes seront fortement représentés dans la population finale. Ce résultat suggère donc que la contribution des Néandertaliens dans le patrimoine génétique des humains modernes est faible voire nulle. Il va à l'encontre de l'hypothèse selon laquelle la trace d'une hybridation importante aurait pu être effacée par dérive génétique. Notre travail souligne l'intérêt de prendre en compte les mouvements des populations au cours du temps, ainsi que leur subdivision. En effet, notre modèle implique un réalisme supplémentaire par rapport à ceux utilisés préalablement, qui ne considéraient que la fusion entre deux populations non-subdivisées et statiques d'un point de vue spatial (Nordborg 1998 ; Serre *et al.* 2004).

5.2.1 Article

{ Page suivante }

A range expansion of modern humans into Europe implies no admixture with Neanderthals

Running title: Absence of admixture between modern humans and Neanderthals

Mathias Currat[1, 2] and Laurent Excoffier [1]

[1]Computational and Molecular Population Genetics Lab, Zoological Institute, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

[2]Laboratoire de Génétique et Biométrie, Département d'Anthropologie et Ecologie, Université de Genève, CP 511, 1211 Genève 24, Switzerland

Abstract

Modern humans (*Homo sapiens sapiens*) displaced Neanderthals in Europe and western Asia between 45,000 and 30,000 BP. Although no Neanderthal lineage is found to date among more than 4,000 mitochondrial DNA sequences in Europe, interbreeding has never been completely excluded. By simulating a range expansion of modern human in Europe from the Near-East, we show that the absence of Neanderthal genes in modern Europeans is compatible with at most 120 admixture events between the two subspecies during 12-15,000 years of coexistence. This very low number strongly suggests that the two populations were indeed not interfertile.

Introduction

The "Neanderthals" (*Homo sapiens neanderthalensis*) constitute a group of hominids, whose particular morphology developed in Europe during the last 350,000 years under the effect of selection and genetic drift, reaching its final form approximately 130,000 years ago[1]. This sub-group of hominids (referred to as *HN* hereafter) populated Europe and western Asia until the arrival of the first modern humans (*Homo sapiens sapiens*, referred to *HS* in the following) approximately 45,000 ago[2]. This arrival apparently drove the Neanderthals to extinction in less than 15,000 years, a replacement process that is still not fully understood[3]. An important question which remains to be assessed is whether Neanderthals could hybridize with modern humans and if they left some traces in the current modern human gene pool. While this hypothesis is excluded under the Recent African Origin model (RAO), which postulates a complete replacement of former members of the genus by *Homo sapiens*, it is central to the tenets of the multiregional hypothesis[4]. From a paleontological and archaeological point of view the debate is still open, even if the supporters of the RAO model[3,5] are gaining momentum over those supporting European regional continuity[6], but see also [7]. Recent

morphological studies support a clear distinction between Neanderthals and modern humans[8], and genetic evidence, such as the clear divergence and monophyly of the *HN* mtDNA control region[9-12], suggested a clear separation of the *HN* and *HS* female lineages[13-16], with a divergence time estimated to lie between 300,000 and 750,000 years ago[9,10]. The complete absence of Neanderthal mtDNA sequences in the current European gene pool, attested from the study of more than 4,000 recorded sequences[17,18,19], strongly supported the absence of Neanderthal mtDNA leakage in the modern gene pool, but it was argued that even if some *HN* genes could have passed in the ancient Cro-Magnon gene pool, it could have been lost through genetic drift[20,21]. Recently, several attempts were made at circumventing the drift problem by the direct sequencing of modern human fossils contemporary with the last Neanderthals[15,19]. Inferred Cro-Magnon sequences were indeed found very similar to those of current Europeans with no trace of Neanderthal influence, even though contamination from modern DNA could not be completely excluded[19]. Under a simple model of instantaneous mixing of Neanderthals and modern humans, a Neanderthal contribution larger than 25% to the modern gene pool could be excluded, but smaller and still significant contributions were found possible[19,22]. Therefore, the problem of the relationships between Neanderthals and modern humans remains fully open.

In order to address this issue, we have developed a more realistic modeling of the range expansion of modern humans in Europe, assumed to be already inhabited by Neanderthals. As will be shown below, we can explain the replacement of Neanderthals by modern humans by a simple competition model between the two species, with a 2.5-4 fold better exploitation of local resources by modern humans as compared to Neanderthals. Moreover, the simulation framework with parameters calibrated by the known duration of the replacement process (about 12,500 years[23]) allows us to estimate the maximum number of fertile admixtures events between the two species compatible with the observed absence of Neanderthal genes in the current gene pool of modern humans.

Results

Modeling the colonization of Europe by modern humans

A digital model of the region encompassing Europe, the Near-East, and North-Africa was build as a matrix of 7,500 cells arranged on a grid. Each of these cells can contain two demes, one occupied by modern humans (*HS*), and the other by Neanderthals (*HN*). At the beginning of the simulation, 1,600 generations ago (corresponding to 40,000 years if a generation time of 25 years is assumed), the *HN* population occupies all the demes corresponding to their estimated range[1] (see Figure 1a). The colonization of Europe by *HS* is then initiated at an arbitrary but plausible point (black arrow on Figure 1a) in western Asia. From this origin, modern humans then progressively colonize neighboring demes, where they were facing competition and potential admixture with Neanderthals (Figure 1). Competition was implemented as a modified Lotka-Volterra model and population interaction was modeled as a density-dependent mating probability with partial fecundity controlling the rate of gene transfer between the two populations. An important and new feature of our model is that the local density of the two populations is logistically regulated, implying that a

newly founded *HS* deme grows logistically until it reaches a given carrying capacity. During this growth period, the *HS* deme can incorporate Neanderthal genes by a density-dependent admixture process. If this happens, a certain fraction of the local *HS* gene pool at equilibrium will consist of genes of *HN* origin, and therefore these Neanderthal genes will have the possibility to be among the *HS* colonizers of new *HS* demes, or to be exchanged with surrounding *HS* demes. Compared to an instantaneous admixture model, we have thus a much more rapid dilution of the modern human gene pool, due to the amplification of the Neanderthal introgressed genes during the logistic growth of the *HS* deme.

Paleodemographic, paleontological and archeological data were used to calibrate the parameters of our competition and admixture model. In our simulations, the replacement of Neanderthals by modern humans in about 500 generations (corresponding to about 15,000 years) is only possible if the carrying capacity of modern humans is larger than that of Neanderthals, which is equivalent to assuming that they had better abilities to exploit local resources, potentially due to their superior technology[1]. Several sets of parameters were found compatible with the known replacement dynamics of the Neanderthal and with available paleodemographic data on Neanderthal and human populations, and six scenarios (A to F) have been studied, as listed in Table 1. The admixture rate, which is the parameter of interest in this study, was allowed to vary and only marginally influenced the cohabitation period and the replacement time of *HN* by *HS* (see Table 1). Note that the cohabitation period at any given place (shown as a narrow black band on Figure 1) is limited to 7-37 generations depending on the scenario listed in Table 1, corresponding to about 175 to 925 years (assuming a generation time of 25 years).

The Neanderthal contribution to the current European gene pool as a function of admixture rates

The expected proportion of Neanderthal genes in the gene pool of modern humans was estimated by coalescent simulations, and is reported in Table 1 for different rates of admixture between Neanderthals and modern humans. At odds with previous estimates[19,22,24], our simulations show that even for very few admixture events, the contribution of the Neanderthal lineages in the current gene pool should be very large (Figure S4). For instance, in scenario A, with a four fold advantage in exploitation of local resource by modern humans, a single fertile admixture event in one deme out of 10 over the whole period of coexistence between *HN* and *HS* should lead to the observation of 38% of *HN* genes in the present mtDNA *HS* gene pool (case A in Table 1). This proportion would be lower but still amount to 15% if the advantage of modern humans was reduced to 1.6 times over Neanderthals with the same admixture rate (case E in Table 1). With higher but still relatively low levels of admixture, a majority of Neanderthal genes should be expected in the current European gene pool (Table 1). For instance, with as much as 2 admixture events per cell over the total coexistence period of Neanderthals and modern humans, more than 95% of the current *HS* gene pool should be tracing back to Neanderthals, for all scenarios with logistic demographic regulation described in Table 1 (scenarios A to F). As shown on Figure 2, the proportion of current lineages that can be traced to Neanderthals is however not uniformly distributed over Europe in case

of interbreeding. A gradient is visible from the source of the range expansion (which shows the largest proportion of modern human genes) towards the margins of the expansion (the British Isles and the Iberic peninsula), which should then be expected to harbor a larger proportion of Neanderthal genes than the rest of Europe (Figure 2). However, this gradient is relatively weak, and the proportion of *HN* lineages at any position is primarily affected by the degree of admixture between the two populations.

Estimation of admixture rates

The present results show that if Neanderthals could freely breed with modern humans, their contribution to our gene pool would be immense. Since no Neanderthal mtDNA sequence has been observed so far among present Europeans, it is of interest to estimate the maximum admixture rate between Neanderthals and modern humans that would be compatible with an absence of Neanderthal genes, accounting for the current sampling effort and drift over the last 30,000 years. Likelihood estimation was performed under a coalescent simulation framework. For each scenario, we estimated the likelihood of different admixture rates from 10,000 coalescent simulations, as reported in Figure 3. Maximum-likelihood estimates are obviously obtained for a total absence of interbreeding between *HS* and *HN*, but here the interest lies in the upper limit of a 95% confidence interval still compatible with an absence of Neanderthal lineages in the European modern gene pool for the different scenarios. We see that the scenarios A to F can be divided into three groups. Scenarios A, C and F lead to very similar upper bounds for the estimation of the maximum admixture rate (~0.015 admixture events per deme, see Table 5). Similarity of results obtained for scenarios A and C show that the fact that the origin of the spread of modern humans was diffused over a large area or concentrated at a single point does not substantially influence our results. Also the implementation of fully symmetric interbreeding between *HN* and *HS* (scenario F) leads to results very similar to those obtained when we only allow breeding between *HN* females and *HS* males (scenario A). The place of origin for modern humans seems more important, as a putative origin in Iran (scenario B) leads to even lower interbreeding rates (~0.01 admixture events per deme) than if the source is located closer to Europe as in scenario A. Finally, scenarios E and D, corresponding to larger carrying capacities of Neanderthals, would be compatible with a larger amount of admixture between the two species (~0.03 admixture events per deme), which is understandable given the longer cohabitation times under these scenarios (21-37 generations) than under scenarios A-C and F (7-12 generations). These estimates can be translated into a maximum number of interbreeding events having occurred over all Europe during the whole replacement process of Neanderthals by modern humans, as reported in Table 1. We find that, depending on the scenario, these maximum estimates range between 34 (scenario B) and 120 (scenario D) admixture events over the whole of Europe , which are extremely low values given the fact that the two populations coexisted for more than 12,000 years in that region.

Discussion

Our simulations show that the mitochondrial evidence in favor of no or very little interbreeding between Neanderthals and modern human is much stronger than previously realized, as it was thought that the current absence of Neanderthal mtDNA genes may have been compatible with a very important contribution of Neanderthal genes (up to 25%) in the gene pool of the early Cro-Magnon populations[19,22]. However, this estimate was based on a very implausible model of evolution, assuming no population subdivision, constant size, and a single and instantaneous admixture event between Neanderthals and modern humans. When a progressive range expansion of modern humans into Europe is modeled, the *maximum* initial input of Neanderthal genes into the Paleolithic European population can thus be estimated to lie between only 0.02% (scenario B) and 0.09% (scenario D) (Table 2). It should be noted that the different scenarios in this study lead to very similar results concerning the expected proportion of Neanderthal genes in our gene pool (Table 1), and the maximum amount of admixture events between Neanderthals and modern humans (Table 1), suggesting that our results are robust to the inherent inaccuracy in the choice of demographic parameters. While we cannot pretend that our model of interaction between Neanderthals and modern humans in Europe accounts for all details of the process, it is considerably more realistic than previous approaches, and it has the advantage of modeling and explaining the observed extinction of Neanderthals in Europe over a period of 12-15,000 years. The large growth rates used in the simulations compensate for the fact that long-range dispersal is not considered in the model, as i increase the speed of the migration wave[25]. These migrations influence the molecular composition of genes in dispersion[25], but they also increase cohabitation times between *HS* and *HS* and consequently their probability of admixture. The simulation of long-range dispersal would thus probably decrease the low admixture rates estimated with our model.

Our finding that even minute amounts of interbreeding between Neanderthals and modern humans should lead to a massive introgression of Neanderthals mtDNAs into the Cro-Magnon gene pool is somehow counter-intuitive and deserves further explanations. The massive introgression process is actually due to both the progressive dilution of the invading gene pool into that of the pre-existing population[26], and to the amplification of introgressed Neanderthal genes during the early stage of the logistic growth of demes at the front of the range expansion. This process bears some resemblance to the success of mutations arising in the wave front of an expanding population[27], but here interbreeding is comparable to a recurrent mutation process. In order to assess the importance of the period of logistic growth compared to a mere dilution process[26], we have modeled a range expansion process where a newly founded deme reaches instantaneously its carrying capacity, and where a given proportion of genes is recruited from the local Neanderthal gene pool. The results of those simulations, also reported in Table 1 and in Figure 1 as case G, show that without logistic growth much larger interbreeding rates would be necessary to have the same impact on current human diversity. Under this scenario, the absence of Neanderthal mtDNA sequences in present Europeans is still compatible with a maximum of about 1,850 fertile breedings between Neanderthal females and Cro-Magnon males, corresponding to a maximum initial input of 1.2% Neanderthal genes into the European Cro-Magnon population. This figure is 20 times larger than when assuming

an initial logistic growth of newly founded populations, but still about 20 times smaller than when assuming a single admixture event and an instantaneous settlement of Europe by modern humans[19]. It implies that the final contribution of the invaded population on the gene pool of the invading population does not only depend on the total amount of gene flow, but also, and to a larger extent, at which time this gene flow occurred.

Introgressed invaders

Another important result of this study is to show that an expanding population or species should have its own genome invaded by that of the invaded population if interbreeding is possible. Interestingly, this phenomenon could explain some documented cases of mitochondrial DNA introgression (e.g.[28,29]). Our model indeed predicts that introgression would occur preferentially in species having gone through a range expansion, and that the introgressing genome would be that of the invaded population and not that of the invasive species. Of course this result should only apply to the part of the genome that is not under selection or that is not linked to the selective advantage of the invaders. If the mitochondrial genome of modern humans was involved in their higher fitness, the absence of observed mtDNA introgression would not necessarily be due to an absence of interbreeding, but would rather result from an active selection process against crosses between Neanderthal females and modern human males, and one would therefore expect to see potential leakage of Neanderthal genes in our nuclear genome. While some evidence for the differential fitness of some mtDNA human genomes in distinct climates has been recently found[30,31], it is unlikely that such differences were involved in the selective advantage of modern humans over Neanderthals. It is indeed doubtful that modern humans coming from the Middle-East would have had mitochondria better adapted to the colder environment of Europe than Neanderthals, who had spent tens of thousands of years in such a climate[1,7]. It is therefore more likely that modern humans' higher technology and higher cognitive abilities[1], resulting in better resource processing and environmental exploitation, have allowed them to out compete Neanderthals, and that mtDNA was selectively neutral in that respect. It should however be kept in mind that our conclusions assume no sex-bias in interbreeding rates. Studies of fossil Y chromosome or nuclear DNA would be needed to examine the basis of this assumption, but it seems difficult to imagine why interbreeding between Neanderthal men and modern human females resulting in the incorporation of Neanderthal genes would have been more frequent than the reverse situation.

Recent range expansions of Neolithic populations

The present approach could certainly be used to model the spread of Neolithic farmers and the extinction of hunter-gatherer practice in Europe in order to get estimates of the contribution of the Paleolithic populations to the current gene pool of present Europeans. While this estimation is beyond the scope of the present paper, the present simulations suggest that even in case of low levels of interbreeding, the Paleolithic gene pool should be at a majority in current European populations. This point is important as it implies that if Neanderthal lineages would have been present among the Paleolithic populations, they should have persisted after the spread of the

Neolithic in Europe. Previous estimations of the Neolithic contribution to the current European genetic pool reach about 50%[32,33]. Assuming that the Neolithic farmers had themselves no Neanderthal component into their gene pool, which is extremely conservative and actually not supported by our simulations (see Figure 2), it implies that our estimates of the initial input of *HN* into the modern pool would have to be roughly multiplied by two, but still be very small (0.07% for scenario A). Note also that the simulation of a pure acculturation process, which amounts to increasing the carrying capacity of populations after the Neolithic by a factor 250 has virtually no effect on the expected proportion of Neanderthal genes in current Europeans for different interbreeding rates (Figure S4). Another argument against a major influence of the Neolithic expansion has been already inferred from mtDNA. European populations present a signal of Paleolithic demographic expansion, which could be dated to about 40 KY ago[34]. The fact that this signal does not date to the Neolithic implies that most of the mtDNA lineages of current Europeans result from a Paleolithic range expansion[35] It is therefore highly likely that the main cause of the current absence of Neanderthal genes in our mtDNA gene pool is their rarity or even absence among the Cro-Magnon population, and not the later replacement of this population by Neolithic farmers, an hypothesis that seems more and more supported by genetic and paleontological data[19].

Even though our model of interaction and competition between Neanderthals and modern humans may still be overly simple compared to reality, it captures two important historical aspects that were neglected in previous studies. The first one is the documented progressive spread of modern humans in Europe (see Figure 1), and the second is the local and progressive demographic growth of Paleolithic populations, with density-dependent interactions with Neanderthals. With these additional sources of realism, our results very strongly supports the view that there was no or only occasional admixture between Neanderthals and modern humans, giving even more credit to the Recent African Origin hypothesis[4,36] to the expense of the multiregional hypothesis of human evolution[37,38], implying a complete replacement of previous members of the *Homo* genus by modern humans.

Methods

Digital map of Europe: The geographical region encompassing Europe, the Middle-East and North-Africa has been modeled as a collection of 7,500 square cells of 2’500 km^2 each, arranged on a two-dimensional grid, whose shape corresponds to the contour of the European and Mediterranean landmass. Each cell harbors two demes, one potentially occupied by modern humans (*Homo sapiens* - *HS*) and one potentially occupied by Neanderthals (*Homo neanderthalensis* - *HN*). Given the estimated range distribution of Neanderthals[1], *HN* demes were allowed in only 3,500 cells, mainly located in the lower part of Europe and in the Near-East (see Figure 1a). Three land bridges have been artificially added to allow the settlement of Great-Britain and Sicily.

Simulation of the colonization of Europe by modern humans: At the beginning of the simulation, 1,600 generations ago (corresponding to 40,000 years ago when assuming a generation time of 25 years), the *HN* demes are all filled at their carying capacity, K_{HN}, and the population *HS* is assumed

to be restricted to a single deme in the Near East at a position corresponding approximately to the present border between Saudi Arabia and Jordan. This source for the spatial and demographic expansion of modern humans into Europe has been choosen arbitrarily, as its exact origin is still debated[23,39,40]. Since we model the evolution of mtDNA, we only simulate the spread of females, but we implicitly assume that there are the same number of males and females in each deme. The source deme for *HS* is assumed to be at its carrying capacity K_{HS} of 40 females, corresponding to a density of about 0.06-0.1 individuals per km^2 (including males and juveniles), in broad agreement with density estimates for Pleistocene hunter-gatherers[41-43]. *HS* individuals can then migrate freely to each of the four neighboring *HS* demes at rate $m/4$. When a *HS* individual enters an empty deme it results in a colonization event, which initiates a local logistic growth process, with intrinsec rate of growth r_{HS} per generation, and with limiting carrying capacity K_{HS}. Interactions between the *HS* and the *HN* demes of the same cell are described below in more detail, and its combination with migrations between *HS* demes results in a wave of advance progressing from the Near-East towards Europe and North Africa. The simulation of such a colonization process has been previously described in absence of competition in a homogeneous square world[35].

Demographic model: We describe here a demographic model of interaction between populations, incorporating competition and interbreeding between individuals of the *HN* and *HS* populations, as well as migration between neighboring demes from the same subdivided population arranged on a 2-dimensional stepping-stone. We distinguish here migrations events between *HN* and *HS* populations from migrations between neighboring *HN* or *HS* populations. We model the former ones as admixture events, whereas the latter ones correspond to true dispersal events. The life cycle of a population at a given generation is as follows: admixture, logistic regulation incorporating competition, followed by migration. This life cycle thus assumes that migration is at the adult stage. In line with previous work[44], the frequency of admixture events is assumed density-dependent. Within a given deme, each of the N_i individuals from the i-th population has a probability $A_{ij} = \gamma_{ij}\left(2N_iN_j\right)/\left(N_i + N_j\right)^2$ to reproduce succesfully with one of the N_j members of the *j*-th population, and γ_{ij} represents the probability that such a mating results in a fertile offspring. Following admixture, population densities are then first updated as $N_i^{'} = N_i\left[1 - A_{ij}\right] + A_{ji}N_j$. Our model of density regulation incorporating competition is based on the Lotka-Volterra interspecific competition model, which is an extension of the logistic growth model[45,46]. For each population, a new density $N_i^{''}$ is calculated from the former density as $N_i^{''} = N_i^{'}\left(1 + r_i(K_i - N_i^{'} - \alpha_{ij}N_j^{'})/K_i\right)$, where r_i is the intrinsic growth rate of the *i*-th population, K_i is its carrying capacity, and α_{ij} is an asymmetric competition coefficient[47, p.274-278]. An α_{ij} value of 1 implies that individuals of the *j*-th population have as much influence on those of population *i* as on their own conspecific, or that competition between populations is as strong as competition within population. Lower values of α_{ij} indicate lower levels of competition between than within populations, a value of zero implying no competition between individuals from different populations. We have decided here not to fix α_{ij} values, but to make them

density-dependent as $\alpha_{ij} = N_j^{'} / (N_i^{'} + N_j^{'})$, reflecting the fact that the influence of the members of a population on the other grows with its density. In the migration phase, each population of each deme can send emigrants to the same population in neighboring demes at rate *m*. $N_i^{''} m$ emigrants are thus sent outwards each generation, and distributed equally among the four neighboring demes, as described previously[35]. If a gene is sent to an occupied deme, the migration event results in gene flow, otherwise it results in the colonization of a new deme. This latter possibility only exists for the population of modern humans, since we assume that Europe was already fully colonized by Neanderthals. Finally, the densities of the two populations are updated as a balance between logistic growth, migration and admixture as $N_i^{'''} = N_i^{''}\left[1-m\right]+I_i$, where I_i is the number of immigrants received from neighboring demes.

Parameter calibration: The speed of the colonization process of modern humans in Europe depends on the parameters of their interaction with Neanderthals, as well as on parameters of migration and logistic growth of the two populations. We have thus calibrated the parameters of our simulation model from available paleodemographic information and from the estimated colonization time of Europe by modern humans. Estimates of the total number of hunter-gatherers living before Neolithic times range between 5 and 10 millions[48-53], of which about 1 million individuals were probably living in Europe. Taking a carrying capacity K_{HS} of 40 females would imply the presence of 220’000 effective mtDNA genes in the 5’500 demes occupied by modern humans in Europe and the Middle-East. Since this number represent only females, the total number of individuals living over has to be at least multiplied by four to include men and juveniles, leading to a total density of about 880’000 *HS* individuals. This value of K_{HS} corresponds to a density of 0.064 individuals per square kilometer, which is close to the value (0.04) used by some previous simulation of modern humans[44,54] and well within the range obtained from actual hunter-gatherer groups (0.01-0.35[55]) or that estimated for ancient hunters-gatherers (0.015-0.2[41-43]). The time required for the colonization of Europe by modern humans is the other information that was used to calibrate the growth rates r_{HS}, the rate of migration m_{HS} and the Neanderthal carrying capacity (K_{HN}), as these three parameters have an influence on the speed of the migration wave[56,57]. Since modern humans arrived in Europe approximately 40' 000 ago and occupied the whole continent by 27’500 BP[23], the colonization process lasted approximately 500 generations, assuming an average generation time of 25 to 30 years[58,59].

Scenarios of modern human range expansion in Europe: Among the many sets of parameter values leading to such a colonization time and the complete disappearance of Neanderthals, we have retained the following scenarios. Scenario A: Origin of *HS* in a single deme of the Near-East at the border between Saoudi Arabia and Jordany, m_{HS} = m_{HN} = 0.25, r_{HN} = 0.4 and K_{HN} = 10, r_{HS} = 0.4, K_{HS} = 40. Note that a value of K_{HN} of 10 corresponds to a total density of about 140’000 Neanderthals over Europe (0.016 individuals per km^2), which is of the same order of magnitude as

the rare available estimates (250,000 Neanderthals[60]). Under this scenario, we have only considered admixture events between *HN* females and *HS* males, such that $\gamma_{HS,HN}$ = 0. Six alternative scenarios have been considered. Scenario B is identical to scenario A, except that the *HS* origin is located in Iran. Scenario C uses the same parameters as scenario A, but the *HS* source is more diffuse and corresponds to a subdivided population of 25 demes surrounding the source deme defined in scenario A. Scenario D is identical to A, but r_{HS} is here equal to 0.8, which is the maximum growth rate estimated for Paleolithic human population[51,61]. Scenario E is identical to A, except that m_{HS} is here much higher and equal to 0.5, implying that 50% of the women are recruited in adjacent demes. The carrying capacity of Neanderthals K_{HN} had to be readjusted for scenarios D and E, which may appear as extreme, in order to maintain a colonization time of about 500 generations. It was indeed set to 25, giving a total density of *HN* of 350,000 individuals over Europe. Scenario F is identical to A, but admixture can occur between *HN* males and *HS* females as well, such that $\gamma_{HS,HN} = \gamma_{HN,HS}$. Finally, scenario G uses the same parameters as A, but a different demographic model. When a cell is colonized by *HS*, it is directly filled at K_{HS} with an initial proportion γ of Neanderthals. Admixture thus occurs when demographic equilibrium is already reached, and not during the demographic growth as in the other models. While the γ values are the true parameters of our model, they may not be very telling per se, and we have therefore chosen to quantify levels of interbreeding between populations using another parameterization, which corresponds to the average number of admixture events per deme between modern humans and Neanderthals. By performing a large series of simulations, the values of γ leading to a given average number of admixture events per deme (e. g. 1/500, 1/100, 1/10, 1, 2, etc...) have been found.

Coalescent simulations: During 1,600 generations, the two populations evolve according to the model described above. *HS* progressively invades the territory of the Neanderthals due to its larger carrying capacity. When demes are reached by the *HS* expansion wave, *HN* and *HS* coexist during several generations in the same cell, and *HN* disappears under the effect of competition. Such a typical demographic transition is shown in Figure S5, together with the amount of admixture between *HN* and *HS* resulting in the integration of Neanderthal genes in the *HS* gene pool. For each scenario and for different values of the amount of interbreeding γ_{ij}, the demography of more than 14,000 demes is thus simulated for 1,600 generations and recorded in a database. The density of all demes, the number of migrants exchanged between demes from the same population, as well as the number of admixture events resulting in gene movements between Neanderthals and modern humans are recorded. This demographic database is then used to simulate the genealogy of samples of 40 genes drawn from 100 demes, representing a total of 4,000 modern human genes distributed over all Europe, and corresponding approximately to the current sampling effort of European mtDNA sequence[17,18,62]. The coalescent simulations proceed as described previously[63]. The average proportion of sampled genes whose ancestors can be traced to some Neanderthal lineages was then computed over 10,000 simulations. The coalescent process allows one to simulate the process of genetic drift, which could have led to the disapearance of Neanderthal

sequences having been integrated in the Cro-Magnon gene pool. The likelihood of each interbreeding coefficient γ_{ij} is thus estimated for the different scenarios by the proportion of 10,000 simulations that lead to a Most Recent Common Ancestor of all 4,000 sampled mtDNA sequences being of modern human origin.

Acknowledgements

Thanks to Nicolas Ray and Pierre Berthier for programming and computing assistance. We are grateful to Monty Slatkin, Arnaud Estoup and Grant Hamilton for their critical reading of the manuscript. This work was supported by a Swiss NSF grant No 3100A0-100800 to LE.

Tables & Figures

Table 4: Expected proportion of Neanderthal lineages in the present modern human gene pool under 7 different demographic scenarios.

Scenario	*Demographic. parameters*[a]				*Cohabitation time*[b]	*Colonization time*[b]	*Expected proportion of HN lineages in modern pool for different rates of admixture*[c]									
	K_{HN}	K_{HS}	r_{HS}	m_{HS}			*1/500*	*1/100*	*1/50*	*1/25*	*1/10*	*1/5*	*1/2*	*1*	*2*	*5*
A	10	40	0.4	0.25	7.5-11.0	400-500	0.00	0.05	0.09	0.15	0.38	0.62	0.91	0.99	1.00	1.00
							0.00	*0.10*	*0.15*	*0.13*	*0.23*	*0.21*	*0.10*	*0.02*	*0.00*	*0.00*
B	10	40	0.4	0.25	7.5-11.0	380-550	0.00	0.04	0.10	0.16	0.50	0.64	0.92	0.99	1.00	1.00
							0.01	*0.08*	*0.15*	*0.17*	*0.24*	*0.18*	*0.08*	*0.02*	*0.00*	*0.00*
C	10	40	0.4	0.25	7.5-11.0	350-500	0.00	0.01	0.05	0.18	0.38	0.57	0.92	0.98	1.00	1.00
							0.02	*0.02*	*0.13*	*0.17*	*0.22*	*0.21*	*0.08*	*0.03*	*0.01*	*0.00*
D	25	40	0.8	0.25	21.5-24.5	430-500	0.00	0.01	0.05	0.05	0.20	0.31	0.68	0.85	0.98	1.00
							0.02	*0.02*	*0.13*	*0.09*	*0.21*	*0.18*	*0.20*	*0.13*	*0.03*	*0.00*
E	25	40	0.4	0.5	31.0-36.5	390-470	0.00	0.01	0.03	0.05	0.15	0.34	0.62	0.86	0.95	1.00
							0.01	*0.03*	*0.05*	*0.09*	*0.14*	*0.21*	*0.17*	*0.11*	*0.05*	*0.01*
F	10	40	0.4	0.25	11.0-12.0	450-510	0.00	0.06	0.05	0.25	0.51	0.62	0.91	0.98	1.00	1.00
							0.01	*0.13*	*0.08*	*0.25*	*0.27*	*0.20*	*0.10*	*0.03*	*0.00*	*0.01*
G	10	40	-	0.25	1.0	140-170	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.05	0.14
							0.00	*0.00*	*0.00*	*0.00*	*0.00*	*0.02*	*0.01*	*0.04*	*0.04*	*0.06*

The expected contribution of Neanderthal lineages in the current gene pool of modern humans (over all the simulated demes) was obtained from 10,000 simulations. Standard deviations are shown in italic. Demographic scenarios: **A**) The basic scenario with realistic parameters **B)** identical to A, with an origin in Iran at the extreme East of the simulated area. **C**) identical to A, but with a diffused source area consisting in 25 demes at K_{HS}=40, instead of only one deme. **D**) identical to A, with r_{HS} = 0.8, and K_{HN} adjusted to 25 in order to obtain realistic colonization times **E**) identical to A, with a migration rate of m_{HS} increased to 0.5 and K_{HN} adjusted to 25 as in D. **F**) identical to A, with interbreeding resulting in symmetrical transfer of genes between modern humans and Neanderthals **G**) identical to A, but with a modified demographic model with carrying capacity K_{HS} being reached instantaneously and with a local recruitment of γK_{HS} Neanderthal lineages. In this scenario, there is thus a single event of admixture at demographic equilibrium and no logistic growth.

[a] K_{HN}: carrying capacity of Neanderthal demes; K_{HS}: carrying capacity of modern human demes; r_{HS}: intrinsic rate of growth of modern human per generation; m_{HS}: migration rate between adjacent modern human demes.

[b] In generation

[c] The different rates of admixture are given in number of admixture events per deme. For instance, a value of 1/10 implies an average of one admixture event for 10 demes for the whole period of cohabitation between Neanderthals and modern humans.

Table 5: Upper limits of the number of admixture events and initial contribution of Neanderthals.

Scenario	*Maximum number of admixture events per deme* [a]	*Maximum number of admixture events over the whole Europe* [b]	*Maximum Neanderthal initial input into modern human gene pool* (%) [c]
A	0.0144	50	0.036
B	0.0097	34	0.024
C	0.0161	56	0.040
D	0.0344	120	0.086
E	0.0282	99	0.070
F	0.0159	56	0.040
G	0.5322	1863	1.330

[a] Upper limit of a 95% confidence interval for the average number of admixture events per deme, which is still compatible with an absence of Neanderthal mtDNA genes in current Europeans.

[b] Upper limit of a 95% confidence interval for the total number of admixture events having occurred over the 3,500 demes in Europe where Neanderthals and modern human have co-existed.

[c] Maximum percentage of Neanderthal input into the initial Paleolithic population. This figure is computed from the previous column by assuming that there were a total of 140,000 reproducing females in the total modern human population in Europe.

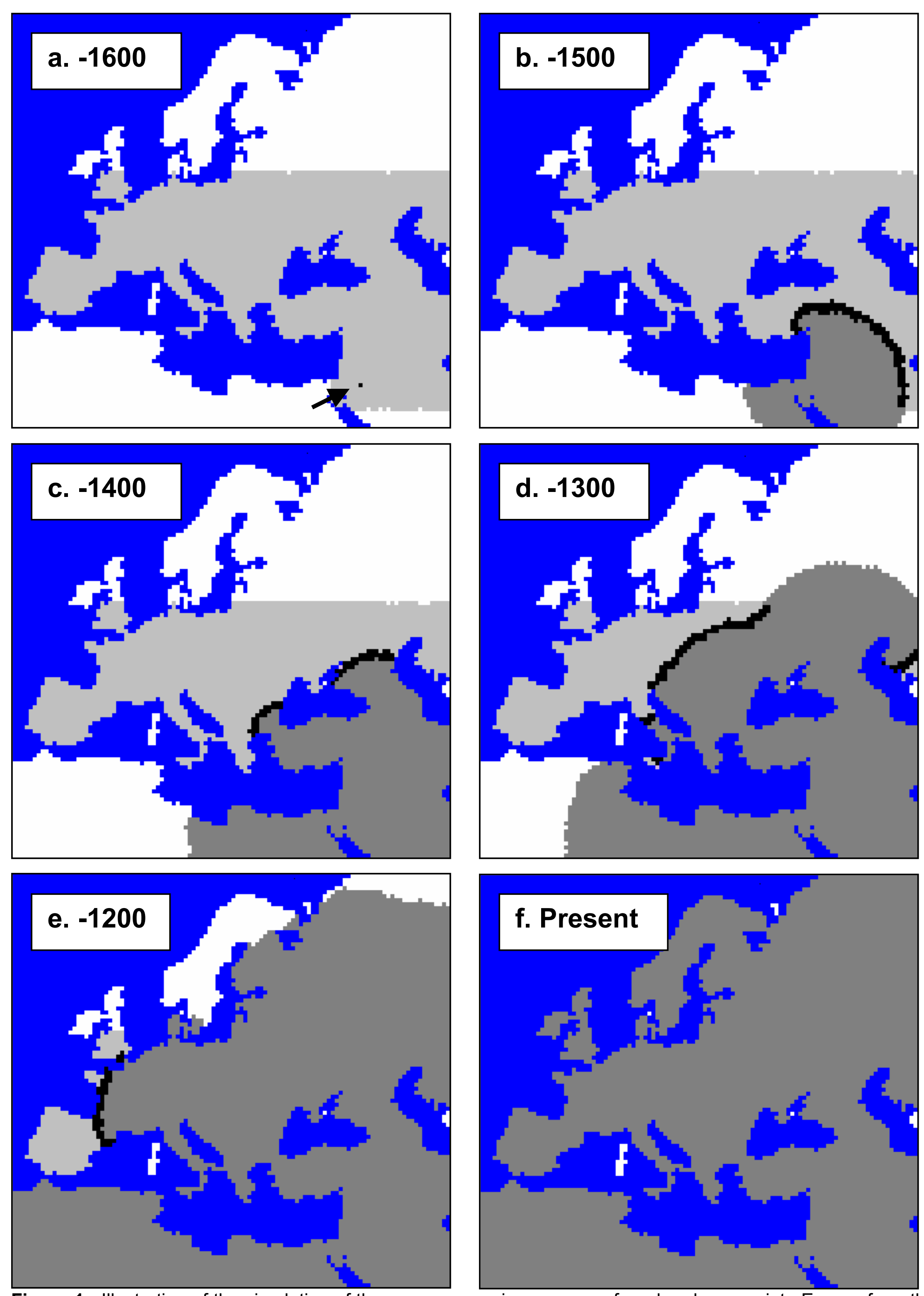


Figure 1 : Illustration of the simulation of the range expansion process of modern humans into Europe from the Near-East, at six different times. Simulations begin 1,600 generations ago, with a part of Europe already colonized by Neanderthals shown as a light grey area, and an origin of modern human expansion indicated by a black arrow (pane a). Panes b-e show the progression of the wave of advance of modern humans into Europe at different times before present. The black band at the front of the expansion wave represents the restricted zone of cohabitation between modern humans and Neanderthals.

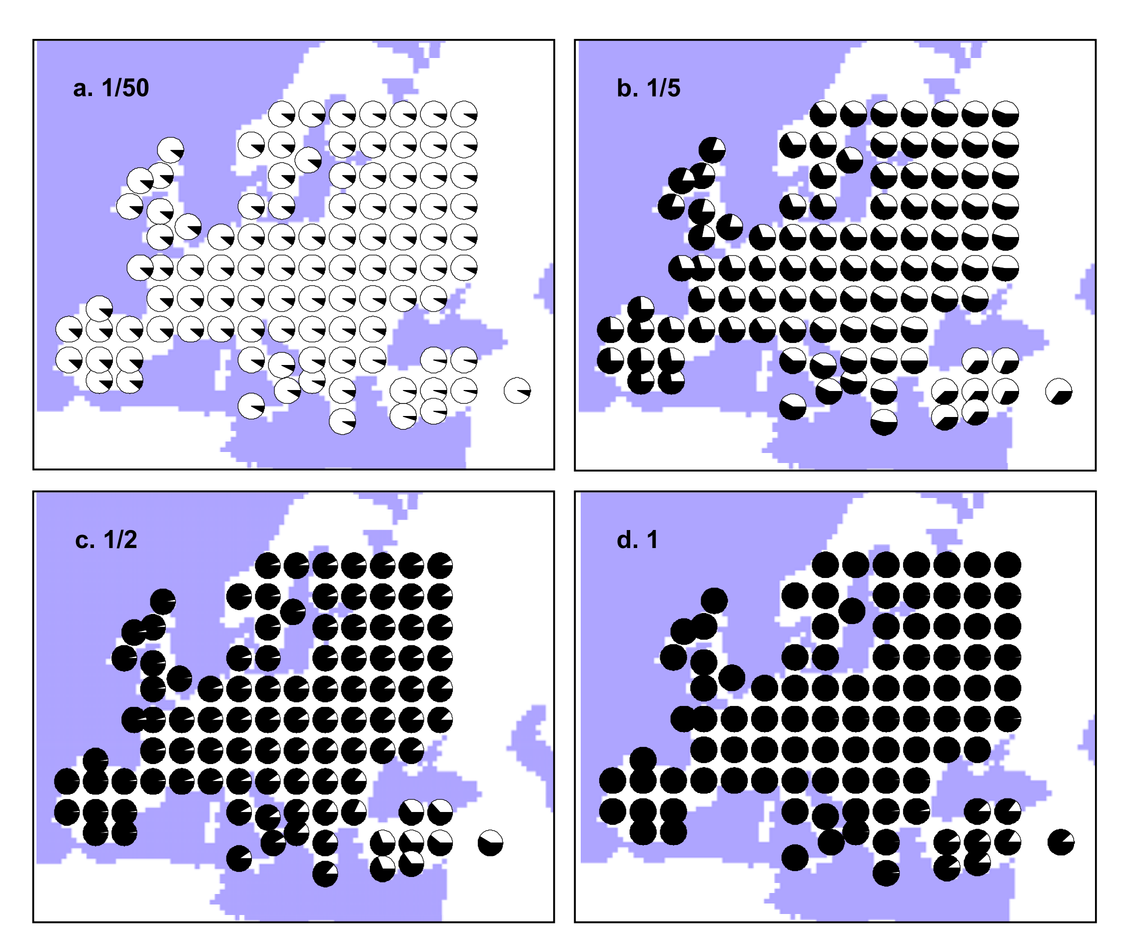


Figure 2: Expected proportion of Neanderthal lineages (in black) among European samples under demographic scenario A (Table 1) at different geographic locations, for different interbreeding rates. **a**. = 1 admixture event on average per 50 demes over the whole period of cohabitation between Neanderthals and modern humans; **b**. = 1 event per 5 demes; **c**. = 1 event per 2 demes; **d**. = 1 event per deme.

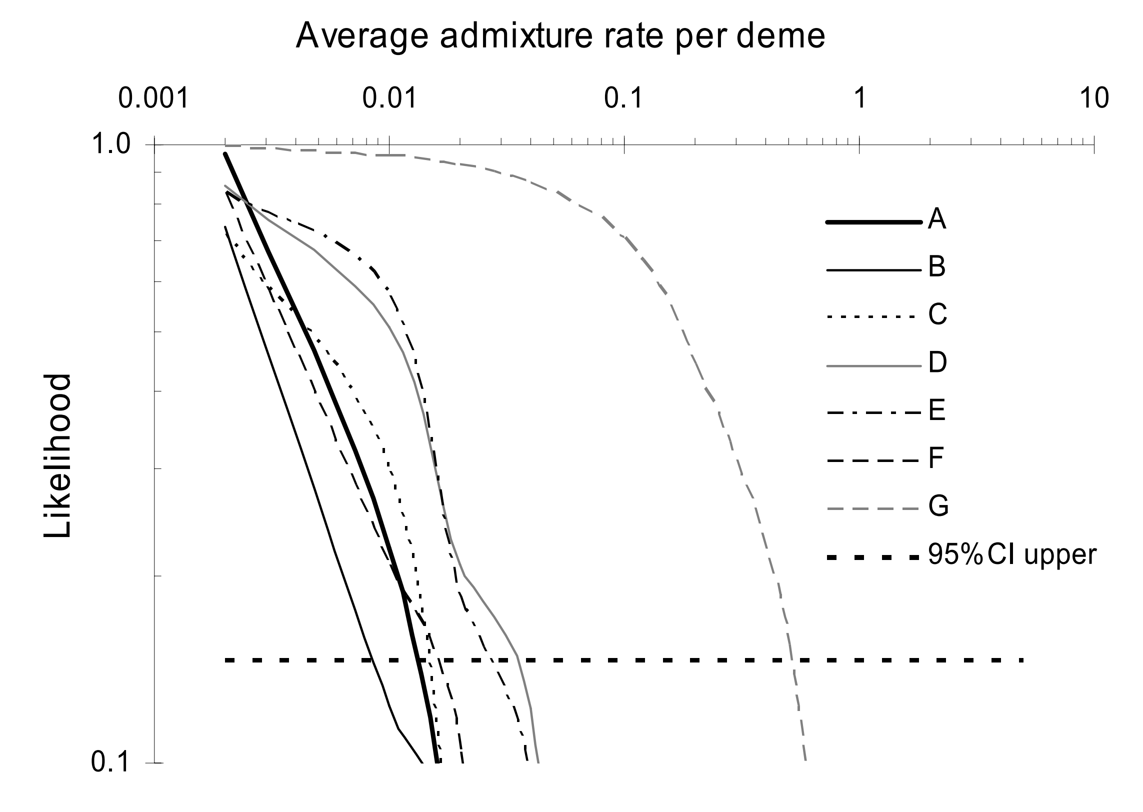


Figure 3: Likelihood of different rates of interbreeding under the seven scenarios described in Table 1. The horizontal bold dashed line corresponds to 14.7% of the maximum likelihood, defining the upper limit of a 95% confidence interval for the interbreeding rates e.g[64].

Additional Figures

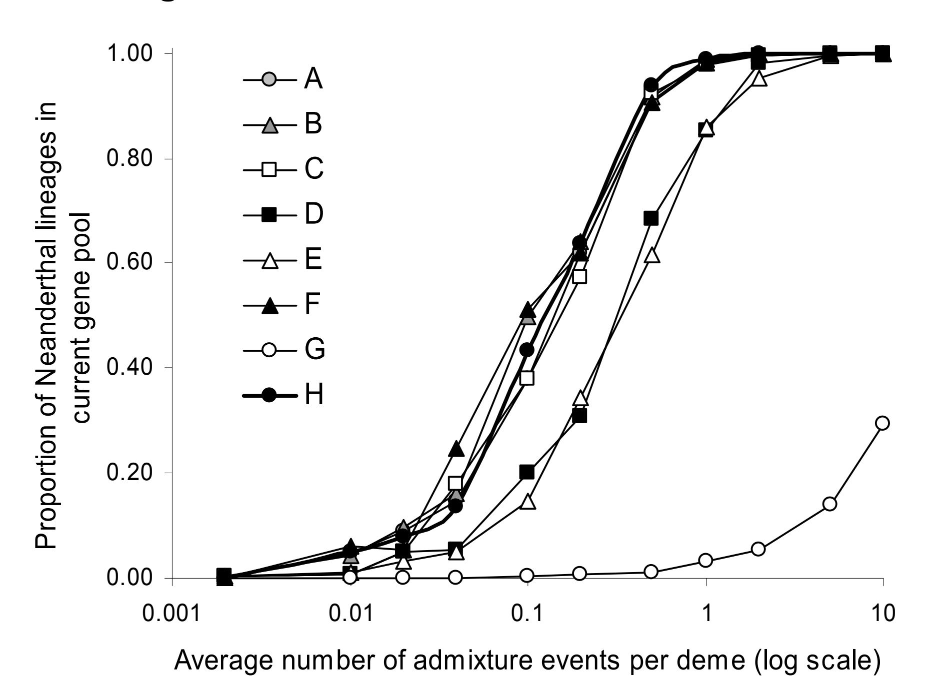


Figure S4: Proportion of Neanderthal lineages in the European population as a function of the average number of admixture events per deme between *HN* and *HS*. These values are given for the seven scenarios mentionned in the article (A - G) and for a new scenario H. This last scenario is similar to A, except that the carrying capacity of the modern humans is increased by a factor 250 at the time of the Neolithic transition (320 generations BP). The influence of this demographic increase on the simulated *HN* proportion is very weak, as shown on this figure.

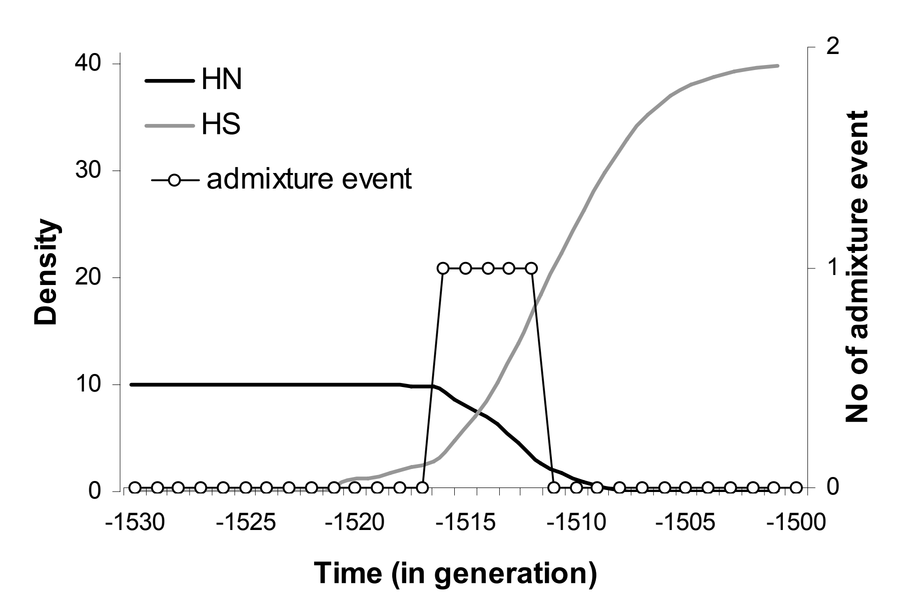


Figure S5: Evolution of the densities of demes *HN* (in black) and *HS* (in gray) within a cell simulated under demographic scenario A for γ_{ij} = 0.4. The cell is colonized by *HS* at time -1520 (0 = present). The thin black line with white circles represents the distribution of admixture events.

References

1. Klein, R.G. Paleoanthropology. Whither the Neanderthals? Science 299, 1525-7 (2003).
2. Mellars, P.A. Archaeology and the population-dispersal hypothesis of modern human origins in Europe. Philos Trans R Soc Lond B Biol Sci 337, 225-34 (1992).
3. Stringer, C. & Davies, W. Archaeology. Those elusive Neanderthals. Nature 413, 791-2 (2001).
4. Excoffier, L. Human demographic history: refining the recent African origin model. Current Opinion in Genetics and Development 12, 675-682 (2002).
5. Rak, Y., Ginzburg, A. & Geffen, E. Does Homo neanderthalensis play a role in modern human ancestry? The mandibular evidence. Am J Phys Anthropol 119, 199-204 (2002).
6. Duarte, C. et al. The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. Proc Natl Acad Sci U S A 96, 7604-9 (1999).
7. Tattersall, I. & Schwartz, J.H. Hominids and hybrids: the place of Neanderthals in human evolution. Proc Natl Acad Sci U S A 96, 7117-7119 (1999).
8. Harvati, K. The Neanderthal taxonomic position: models of intra- and inter-specific craniofacial variation. J Hum Evol 44, 107-32 (2003).
9. Krings, M. et al. Neandertal DNA sequences and the origin of modern humans. Cell 90, 19-30 (1997).
10. Krings, M., Geisert, H., Schmitz, R.W., Krainitzki, H. & Paabo, S. DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. Proc Natl Acad Sci U S A 96, 5581-5 (1999).
11. Ovchinnikov, I.V. et al. Molecular analysis of Neanderthal DNA from the nothern Caucasus. Nature 404, 490-493 (2000).
12. Cooper, A. et al. Human origins and ancient human DNA. Science 292, 1655-6 (2001).
13. Scholz, M. et al. Genomic differentiation of Neanderthals and Anatomically modern man allows a fossil-DNA-based classification of morphologically indistinguishable hominid bones. Am J Hum Genet 66, 1927-1932 (2000).
14. Krings, M. et al. A view of Neandertal genetic diversity. Nat Genet 26, 144-6 (2000).
15. Caramelli, D. et al. Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proc Natl Acad Sci U S A 100, 6593-7 (2003).
16. Schmitz, R.W. et al. The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. Proc Natl Acad Sci U S A 99, 13342-7 (2002).
17. Sykes, B. The molecular genetics of European ancestry. Philos Trans R Soc Lond B Biol Sci 354, 131-8; discussion 138-9. (1999).
18. Handt, O., Meyer, S. & von Haeseler, A. Compilation of human mtDNA control region sequences. Nucleic Acids Research 26, 126-12 (1998).
19. Serre, D. et al. No Evidence of Neandertal mtDNA Contribution to Early Modern Humans. PLoS Biol 2, E57 (2004).
20. Relethford, J.H. Absence of regional affinities of Neandertal DNA with living humans does not reject multiregional evolution. Am J Phys Anthropol 115, 95-8 (2001).
21. Hagelberg, E. Recombination or mutation rate heterogeneity? Implications for Mitochondrial Eve. Trends Genet 19, 84-90 (2003).
22. Nordborg, M. On the probability of Neanderthal ancestry. Am J Hum Genet 63, 1237-40 (1998).
23. Bocquet-Appel, J.-P. & Demars, P.Y. Neanderthal contraction and modern human colonization of Europe. Antiquity 74, 544-552 (2000).
24. Gutierrez, G., Sanchez, D. & Marin, A. A reanalysis of the ancient mitochondrial DNA sequences recovered from Neandertal bones. Mol Biol Evol 19, 1359-66 (2002).
25. Nichols, R.A. & Hewitt, G.M. The genetic consequences of long distance dispersal during colonization. Heredity 72, 312-317 (1994).
26. Chikhi, L., Nichols, R.A., Barbujani, G. & Beaumont, M.A. Y genetic data support the Neolithic demic diffusion model. PNAS 99, 11008-11013 (2002).
27. Edmonds, C.A., Lillie, A.S. & Cavalli-Sforza, L.L. Mutations arising in the wave front of an expanding population. PNAS 101, 975-979 (2004).
28. Bernatchez, L., Glémet, H., Wilson, C.C. & Danzmann, R.G. Introgression and fixation of Arctic char (Salvelinus alpinus) mitochondrial genome in an allopatric population of brook trout (Salvelinus fontinalis). Canadian Journal of Fisheries and Aquatic Science 52, 179-185 (1995).
29. Shaw, K.L. Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: What mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. Proc Natl Acad Sci U S A 99, 16122-16127 (2002).
30. Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V. & Wallace, D.C. Effects of Purifying and Adaptive Selection on Regional Variation in Human mtDNA. Science 303, 223-226 (2004).
31. Mishmar, D. et al. Natural selection shaped regional mtDNA variation in humans. PNAS 100, 171-176 (2003).
32. Chikhi, L. Admixture and the demic diffusion model in Europe. in Examining the farming/language dispersal hypothesis (eds. Bellwood, P. & Renfrew, C.) 435-447 (McDonald Institute Monographs, Cambrigs, 2002).

33. Barbujani, G. & Dupanloup, I. DNA Variation in Europe: estimating the demographic impact of Neolithic dispersals. in Examining the farming/language dispersal hypothesis (eds. Bellwood, P. & Renfrew, C.) 421-431 (McDonald Institute Monographs, Cambrigs, 2002).
34. Excoffier, L. & Schneider, S. Why hunter-gatherer populations do not show sign of Pleistocene demographic expansions. Proceedings of the National Academy of Sciences USA 96, 10597-10602 (1999).
35. Ray, N., Currat, M. & Excoffier, L. Intra-deme molecular diversity in spatially expanding populations. Mol Biol Evol 20, 76-86 (2003).
36. Stringer, C. Modern human origins: progress and prospects. Philos Trans R Soc Lond B Biol Sci 357, 563-79 (2002).
37. Eckhardt, R.B., Wolpoff, M.H. & Thorne, A.G. Multiregional evolution. Science 262, 973-4 (1993).
38. Wolpoff, M.H., Hawks, J. & Caspari, R. Multiregional, not multiple origins. Am J Phys Anthropol 112, 129-36 (2000).
39. Djindjian, F., Koslowski, J. & Otte, M. Le Paléolithique supérieur en Europe, 474 (Armand Colin, Paris, 1999).
40. Kozlowski, J. & Otte, M. The formation of the Aurignacian. Journal of Anthropological Research 56, 513-524 (2000).
41. Alroy, J. A multispecies overkill simulation of the end-Pleistocene megafaunal mass extinction. Science 292, 1893-1896 (2001).
42. Bocquet-Appel, J.-P. & Demars, P.Y. Population Kinetics in the Upper Palaeolithic in western Europe. Journal of Archaeological Science 27, 551-570 (2000).
43. Steele, J., Adams, J.M. & Sluckin, T. Modeling Paleoindian dispersals. World Archeology 30, 286-305 (1998).
44. Barbujani, G., Sokal, R.R. & Oden, N.L. Indo-European origins: a computer-simulation test of five hypotheses. Am J Phys Anthropol 96, 109-32. (1995).
45. Lotka, A.J. The growth of mixed populations: two species competing for a common food supply. Journal of the Washington academy of Sciences 22, 461-469 (1932).
46. Volterra, V. Variations and fluctuations of the numbers of individuals in animal species living together (Reprinted in 1931). in Animal Ecology (ed. Chapman, R.N.) (Mc Graw Hill, New York, 1926).
47. Begon, M., Harper, J.L. & Townsend, C.R. Ecology, 1068 (Blackwell Science, Oxford, 1996).
48. Lee, R.B. & DeVore, I. Problems in the study of hunters and gatherers. in Man the hunter (eds. Lee, R.B. & DeVore, I.) 4-12 (Aldine Publishing Company, Chicago, 1968).
49. Coale, A.J. The history of the human population. Scientific American 231, 40-51 (1974).
50. Hassan, F.A. The peopling of the World. in Demographic archaeology 193-208 (Academic Press, New York, 1981).
51. Ammerman, A. & Cavalli-Sforza, L.L. The Neolithic transition and the genetics of populations in Europe, 176 (Princeton University Press, Princeton, New Jersey, 1984).
52. Weiss, K.M. On the number of members of the Genus Homo who have ever lived, and some evolutionary implications. Human Biology 56, 637-649 (1984).
53. Landers, J. Reconstructing ancient populations. in The Cambridge Encyclopedia of Human Evolution. (eds. Jones, S., Martin, R. & Pilbeam, D.) 402-405 (Cambrige University Press, London, 1992).
54. Rendine, S., Piazza, A. & Cavalli-Sforza, L. Simulation and separation by principal components of multiple demic expansions in Europe. Am. Nat. 128, 681-706 (1986).
55. Binford, L.R. Constructing frames of reference. An analytical method for archaeological theory building using hunter-gatherer and environmental data sets, 563 (University of California Press, Berkeley, 2001).
56. Skellam, J.G. Random dispersal in theoretical populations. Biometrica 38, 196-218 (1951).
57. Fisher, R.A. The wave of advance of advantageous genes. Annals of Eugenics 7, 355-369 (1937).
58. Helgason, A., Hrafnkelsson, B., Gulcher, J.R., Ward, R. & Stefansson, K. A Populationwide Coalescent Analysis of Icelandic Matrilineal and Patrilineal Genealogies: Evidence for a faster Evolutionary Rate of mtDNA Lineages than Y Chromosomes. Am J Hum Genet 72, 00-00 (2003).
59. Tremblay, M. & Vezina, H. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. American Journal of Human Genetics 66, 651-8 (2000).
60. Biraben, J.-N. L'évolution du nombre des hommes. Population et Sociétés 394, 1-4 (2003).
61. Young, D.A. & Bettinger, R.L. Simulating the global human expansion in the late pleistocene. Journal of Archaeological Science 22, 89-92 (1995).
62. Richards, M. et al. Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 59, 185-203. (1996).
63. Currat, M., Ray, N. & Excoffier, L. SPLATCHE: A program to simulate genetic diversity taking into account environmental heterogeneity. Molecular Ecology Notes 4, 139-142 (2004).
64. Kalbfleisch, J.G. Probability and Statistical Inference, 360 (Springer Verlag, New York, 1985).

6 Expansion des populations néolithiques en Europe

6.1 Introduction

La structure génétique des populations européennes est modifiée par les nombreux facteurs dont nous avons déjà parlé en introduction du chapitre 2, et elle peut potentiellement porter la trace d'événements démographiques majeurs (Menozzi *et al.* 1978 ; Sokal 1991b ; Barbujani *et al.* 1995 ; Barbujani et Bertorelle 2001). Si la position géographique des populations européennes[1] ainsi que les barrières géographiques qui les séparent jouent vraisemblablement un rôle plus important que leurs affinités linguistiques (Sokal 1988 ; Sokal *et al.* 1988 ; Barbujani et Sokal 1991 ; Barbujani *et al.* 1996 ; voir aussi Hurles *et al.* 1999 ; Rosser *et al.* 2000 ; Bosch *et al.* 2001 ; Stefan *et al.* 2001 ; Brion *et al.* 2003), il semble tout de même qu'à l'échelle continentale elles portent certaines traces génétiques d'une histoire démographique commune :

- Les populations européennes sont génétiquement très homogènes, particulièrement en ce qui concerne le génome mitochondrial (Horai et Hayasaka 1990 ; Jorde *et al.* 1995 ; Comas *et al.* 1997), mais également pour le chromosome Y puisque l'Europe est le continent dont le F_{ST}[2] est le plus faible (Roewer *et al.* 2000 ; Hammer *et al.* 2001 ; Kayser *et al.* 2001). Cette homogénéité importante a déjà été observée à l'aide des marqueurs classiques[3] (Cavalli-Sforza et Piazza 1993 ; Dugoujon *et al.* 2004) et a été interprétée comme étant le résultat d'une origine commune récente des populations européennes (Pult *et al.* 1994).

- L'hypothèse d'une origine commune récente a été renforcée par l'observation de la trace d'une expansion démographique paléolithique dans le génome des populations européennes (Bertranpetit *et al.* 1995 ; Calafell *et al.* 1996 ; Comas *et al.* 1996 ; Francalacci *et al.* 1996 ; Comas *et al.* 1997 ; Excoffier et Schneider 1999 ; Pritchard *et al.* 1999 ; Shen *et al.* 2000).

- Des gradients de fréquences alléliques entre le Proche-Orient et le nord-ouest de l'Europe, que nous appelerons "SE-NO" pour <u>S</u>ud/<u>E</u>st – <u>N</u>ord/<u>O</u>uest, ont été régulièrement observés à l'aide de différents systèmes génétiques. Environ 1/3 des marqueurs classiques montrent de tels gradients (Menozzi *et al.* 1978 ; Sokal et Menozzi 1982 ; Sokal *et al.* 1991 ; Barbujani et Pilastro 1993 ; Cavalli-Sforza *et al.* 1994 ; Piazza *et al.* 1995 ; Simoni *et al.* 1999), qui ont également été observés à l'aide de marqueurs nucléaires moléculaires (Chikhi *et al.* 1998) et, de façon un peu moins évidente, par certains polymorphismes du chromosome Y (Semino *et al.* 1996 ; Malaspina *et*

---

[1] Le modèle d'"isolation par la distance" (Malécot 1948, 1955) propose que les migrations sont les principales responsables de la structure génétique des populations humaines et que, plus ces dernières sont éloignées géographiquement, plus elles vont être différenciées génétiquement. Ce modèle permet d'expliquer une grande partie de la variation génétique des populations humaines (Morton 1977 ; Morton 1982), principalement à l'échelle continentale ou mondiale (p. ex. : Sanchez-Mazas *et al.* 1994 ; Poloni *et al.* 1995).

[2] La statistique F_{ST} (Wright 1943) traduit le degré de différenciation génétique entre groupes d'individus et peut être calculée de différentes manières, soit entre deux populations (Reynolds *et al.* 1983 ; Slatkin 1991, 1995), soit de manière globale entre toutes les populations (Cockerham 1969 ; 1973 ; Excoffier *et al.* 1992).

[3] Les marqueurs que l'on appelle couramment "classiques" sont constitués par des locus situés sur les systèmes sanguins (Rhésus, ABO, etc…), immunitaires (HLA) ou par des protéines ou allozymes (GF, HP, CP, etc…). Voir Mourant (1976), Roychoudhury (1988) ou Tills (1983) pour plus de détails.

al. 1998 ; Casalotti *et al.* 1999 ; Quintana-Murci *et al.* 1999 ; Barbujani et Chikhi 2000 ; Hill *et al.* 2000b ; Rosser *et al.* 2000). Jusqu'à récemment (Richards *et al.* 2002), aucun gradient de fréquence équivalent n'avait été observé pour le génome mitochondrial (Richards *et al.* 1996 ; Richards *et al.* 1998 ; Richards *et al.* 2000), excepté le long de la Méditerranée (Simoni *et al.* 2000). La présence de ces gradients a été principalement attribuée à la diffusion des premiers agriculteurs originaires du Proche-Orient (Menozzi *et al.* 1978 ; Ammerman et Cavalli-Sforza 1984 ; Casalotti *et al.* 1999 ; Barbujani et Chikhi 2000 ; Hill *et al.* 2000b ; Rosser *et al.* 2000 ; Quintana-Murci *et al.* 2003). Plus récemment, des auteurs ont mis en doute cette interprétation, en proposant que ces gradients résultent soit de l'arrivée des premiers Hommes modernes (Richards *et al.* 1996 ; Barbujani *et al.* 1998 ; Barbujani et Bertorelle 2001 ; Barbujani et Dupanloup 2002), soit d'un effet sélectif (Fix 1996). Cette sélection pourrait avoir eu lieu sur des gènes de résistance aux maladies infectieuses chez l'Homme, maladies transmises par des pathogènes venant des animaux. Plus les agriculteurs ont été près du lieu d'origine de la domestication, et plus longtemps ils auraient subi cette sélection (Fix 1996). L'observation de gradients de fréquences pour plusieurs locus différents est cependant un argument contre la sélection, dont les effets se font généralement sentir sur un nombre restreint de locus. Il faut noter qu'un gradient "SE-NO" pourrait également être expliqué par la différence entre deux anciennes populations situées à chacune de ses extrémités.

La signature génétique commune à toutes les populations européennes pourrait donc être le résultat d'une expansion démographique récente (< 100'000 ans) à partir de l'est du continent, vraisemblablement du Proche-Orient. Deux événements démographiques principaux peuvent potentiellement avoir généré cette signature. Il s'agit, premièrement, de l'arrivée des premiers *Homo sapiens sapiens* en Europe il y a environ 45'000 ans (Otte 2000), et, deuxièmement, de la transition Néolithique, qui a débuté au Proche-Orient il y a plus de 10'000 ans (Harris 1996 ; Whittle 1996 ; Thorpe 1999 ; Mazurié de Keroualin 2003).

Ce sont donc à ces deux événements que nous nous intéressons dans ce chapitre à l'aide de l'approche par simulation présentée dans le chapitre 4. Nous voulons savoir quels scénarios démographiques, simulés lors de ces phases de peuplement, peuvent être associés avec la structure génétique des populations actuelles. Nous présentons ci-dessous un article soumis à publication (section 6.2.1).

6.2 Diversité génétique en Europe après le Néolithique

Le Néolithique correspond à une période de transition extrêmement importante dans la préhistoire humaine : le passage de la phase de collecte de nourriture à la phase de production de nourriture. Ce changement a également profondément affecté la culture matérielle des populations, mais nous ne tenons compte ici que de l'aspect économique du Néolithique européen.

Il est maintenant acquis que le Néolithique n'est pas endémique au continent européen, mais qu'il a été importé depuis le Proche-Orient, via l'Anatolie. La mise en place des techniques

agropastorales au Proche-Orient et en Anatolie n’a pas été une transition rapide puisqu’elle a duré près de 4'000 ans (11’000-6'800 BC, Mazurié de Keroualin 2001, voir aussi Appenzeller *et al.* 1998). La diffusion de ces techniques jusqu’aux marges ouest et nord de l’Europe s’est faite en moins de 4'000 ans (6'800 - 3'000 BC, Mazurié de Keroualin 2001). Il faudra cependant près d’un millénaire supplémentaire pour que les régions périphériques, notamment la Scandinavie et l’est de l’Europe, adoptent complètement l’économie agropastorale (Zvelebil et Zvelebil 1988 ; Arias 1999).

Deux modèles extrêmes ont été utilisés pour décrire la diffusion du Néolithique en Europe :

i) Le modèle de diffusion démique (DDM : Clark 1965; Ammerman et Cavalli-Sforza 1971).

ii) Le modèle de diffusion culturelle ou acculturation (CDM : Zvelebil et Zvelebil 1988).

Le modèle de diffusion démique implique que les premiers agriculteurs ont connu une forte croissance démographique et qu’ils ont ainsi été forcés de migrer dans les régions voisines, emmenant leurs nouvelles technologies avec eux. Ils auraient ainsi colonisé l’Europe entière par migrations successives, de proche en proche. Ce modèle peut être décrit comme une expansion spatiale et démographique de la population néolithique à l’échelle continentale, avec une contribution minime, sinon nulle, des populations de chasseurs-collecteurs autochtones. D’un point de vue génétique, ce déplacement de population aurait entraîné la diffusion des gènes des premiers agriculteurs proche-orientaux dans toute l’Europe. La proportion des gènes "proche-orientaux" diminuerait progressivement en direction du nord-ouest de l’Europe, au fur et à mesure de l’incorporation de gènes "indigènes". Dans le cas le plus extrême d’un remplacement complet des chasseurs-collecteurs, leurs gènes auraient été effacés du patrimoine génétique européen.

A l’opposé, le processus d’acculturation n’entraîne pas de mouvement de gènes, mais implique seulement la transmission des connaissances agropastorales de proche en proche. Selon ce second modèle, la structure génétique prénéolithique n’aurait peu (ou pas) été influencée par la diffusion des techniques agropastorales.

Les recherches archéologiques ont montré qu’aucun de ces deux modèles à lui seul ne permet d’expliquer la diffusion du Néolithique sur l’ensemble du continent européen (Zvelebil 1986 ; Arias 1999 ; Gronenborg 1999 ; Mazurié de Keroualin 2001). Il semble en effet très improbable qu’un mouvement de population se soit fait depuis le Proche-Orient jusque dans les régions périphériques de l’Europe (Pinhasi *et al.* 2000 ; Zvelebil 2000). La diffusion des techniques agropastorales s’est faite par une succession d’événements d’acculturation dans certaines régions et de remplacement de populations dans d’autres (Mazurié de Keroualin 2001; Gallay 2004).

L’importance du Néolithique sur la structure génétique européenne actuelle dépend principalement de son mode de diffusion. Théoriquement, les données génétiques devraient donc permettre de savoir lequel de ces deux types d’événements (acculturation ou remplacement) a été le plus important et d’estimer ainsi l’influence exacte du Néolithique dans le patrimoine génétique européen. Cependant, la contribution génétique des premiers agriculteurs au patrimoine européen

actuel est très controversée et dépend fortement du type de système étudié, ainsi que du type de méthodologie utilisée (Richards *et al.* 1996 ; Cavalli-Sforza et Minch 1997 ; Richards *et al.* 1998 ; Sykes 1999 ; Richards *et al.* 2000 ; Semino *et al.* 2000a ; Barbujani et Bertorelle 2001 ; Torroni *et al.* 2001 ; Barbujani et Dupanloup 2002 ; Chikhi *et al.* 2002 ; Richards 2003). Actuellement, les estimations les plus contradictoires font état, soit d'une contribution néolithique proche, mais supérieure à 50% (Barbujani et Dupanloup 2002 ; Chikhi 2002), soit d'une contribution inférieure à 25% (Richards 2003).

Nous avons choisi de simuler les deux modèles proposés (DDM et CDM), ainsi qu'une série de scénarios intermédiaires pour lesquels la contribution des chasseurs-collecteurs à la constitution de la population néolithique varie. Nous avons simulé l'arrivée des premiers Hommes modernes en Europe de la même manière que dans la recherche du chapitre précédent (5), suivie de la diffusion du Néolithique à partir de la zone d'origine de l'élevage et de l'agriculture au Levant (Lev-Yadun *et al.* 2000). Ce modèle inclut à la fois de la compétition et des échanges génétiques entre chasseurs-collecteurs et agriculteurs. Un flux génique est simulé uniquement depuis les chasseurs-collecteurs vers les agriculteurs, symbolisant les conséquences soit des mariages mixtes, soit de l'acculturation (adoption des techniques néolithiques par les chasseurs-collecteurs). Ces échanges génétiques ne peuvent avoir lieu que pendant la phase de cohabitation entre les deux populations, puisque les chasseurs-collecteurs disparaissent après quelques générations de contact avec les agriculteurs sous l'effet de la compétition.

Nos simulations constituent la première démonstration formelle que des gradients de fréquences alléliques peuvent être générés par l'expansion paléolithique des premiers européens modernes et qu'ils sont quasiment indépendants de la contribution de ceux-ci au patrimoine génétique néolithique. Contrairement à ce qui avait été proposé préalablement (Ammerman et Cavalli-Sforza 1984 ; Rendine *et al.* 1986 ; Barbujani *et al.* 1995), un remplacement complet ou presque complet des chasseurs-collecteurs pendant le Néolithique n'est pas nécessaire à la présence de ces gradients. Par contre, l'observation de ces gradients est fortement dépendante de l'âge des mutations étudiées, et par conséquent du type de données moléculaires utilisé. Nos simulations montrent en effet que plus une mutation est "ancienne" et plus sa probabilité d'être distribuée sous forme de gradient le long de l'axe de diffusion d'une expansion augmente. Cette observation explique pourquoi le génome mitochondrial ne permet pas l'observation de gradients, alors que les autres systèmes (chromosome Y, nucléaires) le peuvent. En effet, les marqueurs moléculaires typés sur le chromosome Y et les autosomes (SNPs et STRs) ainsi que les marqueurs "classiques" sont sujets à un important biais de recrutement[1] (Rogers et Jorde 1996), auquel échappe le génome mitochondrial, qui est principalement étudié sur la base de séquences d'ADN complètes. Ce biais de recrutement provoque la sous-représentation des mutations "récentes" dans

---

[1] Voir la page 182 pour une définition du "biais de recrutement", ou "ascertainment bias" en anglais.

les données, et par conséquent une augmentation de la proportion de mutations qui montrent des gradients génétiques.

Nous avons également montré dans cette étude que les distributions "mismatch" unimodales montrées par la presque totalité des populations européennes (Di Rienzo et Wilson 1991 ; Bertranpetit *et al.* 1995 ; Calafell *et al.* 1996 ; Comas *et al.* 1996 ; Corte-Real *et al.* 1996 ; Francalacci *et al.* 1996 ; Malyarchuk et Derenko 2001 ; Nasidze et Stoneking 2001), excepté les Saamis (Sajantila *et al.* 1995), sont compatibles avec une forte contribution paléolithique dans le patrimoine mitochondrial. Notre méthode ne permet cependant pas d'estimer précisément cette contribution. En revanche, il n'est pas possible de tirer une quelconque conclusion sur la contribution des chasseurs-collecteurs dans la lignée masculine sur la base des distributions "mismatch" calculées à l'aide de SNPs localisés sur le chromosome Y. En effet, nos simulations montrent que les distributions "mismatch" tirées de données pour lesquelles il existe un biais de recrutement sont majoritairement multimodales, quel que soit le scénario démographique simulé. Le biais de recrutement a donc tendance à effacer la signature des événements démographiques dans les distributions "mismatch".

6.2.1 Article

{ Page suivante }

The effect of the Neolithic expansion on European molecular diversity

Running title: SNP diversity in Europe after a range expansion

Mathias Currat[1,2] and Laurent Excoffier [2]

[1]Laboratoire de Génétique et Biométrie, Département d'Anthropologie et Ecologie, Université de Genève, CP 511, 1211 Genève 24, Switzerland

[2]Computational and Molecular Population Genetics Lab, Zoological Institute, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

Abstract

We performed extensive and realistic simulations of the colonization process of Europe by Neolithic farmers, as well as their potential admixture and competition with local Paleolithic hunter-gatherers. We find that minute amounts of gene flow between Paleolithic and Neolithic populations should lead to a massive Paleolithic contribution to the current gene pool of Europeans. This large Paleolithic contribution is not expected under the demic diffusion model, which postulates that agriculture diffused over Europe by a massive migration of individuals from the Near-East. However, genetic evidence in favor of this model mainly consisted in the observation of allele frequency clines over Europe, which are shown here to be equally likely under a pure demic diffusion or a pure acculturation model. The examination of the consequence of range expansions on SNP diversity reveals that an ascertainment bias consisting in selecting SNPs with high frequencies will promote the observation of genetic clines (which are not expected for random SNPs) and will lead to multimodal mismatch distributions. We conclude that the different patterns of molecular diversity observed for Y chromosome and mtDNA can be at least partly due to an ascertainment bias when selecting Y chromosome SNPs for studying European populations.

Introduction

Two opposing scenarios have been invoked to account for the spread of agriculture in Europe. The demic diffusion (*DD*) model assumes that the Neolithic transition diffused in Europe from the Middle East by an important movement of population (Ammerman & Cavalli-Sforza 1984), without much contact with local Paleolithic populations. On the contrary, the Cultural Diffusion (*CD*) model assumes that the Neolithic transition occurred mainly through the transmission of agricultural techniques (Zvelebil & Zvelebil 1988) without large movements of populations. Archaeological evidence suggests that the dynamics of the spread of agriculture over Europe has been complex,

with a succession of migration phases and local admixture (e.g. Arias 1999; Gronenborg 1999; Mazurié de Keroualin 2003; Zvelebil 1986).

Genetic evidence has been inconclusive so far on the amount of Paleolithic lineage incorporated into the current European gene pool, despite a considerable amount of genetic data available on European populations. This is disappointing since the *DD* and the *CD* models lead to quite different predictions concerning the amount of the current European gene pool tracing back to Paleolithic or Neolithic populations. Under the *CD* model, the current genetic pool should mainly results from hunter-gatherers lineages, while the Near-East Neolithic lineages should be prevalent in the European genetic pool under the *DD* model. The Neolithic contribution to the current European gene pool has been estimated using various approaches, and has led to contradicting results. Depending on the markers used and the type of analyses performed, it varies from a Neolithic contribution smaller than 25% (Richards 2003), to values larger than 50% (Barbujani & Dupanloup 2002; Chikhi 2002).

The analysis of classical nuclear markers and Y chromosomes has also often revealed the presence of allele frequency clines (*AFC*) along a South-East to North-West axis (Barbujani & Pilastro 1993; Chikhi et al. 1998; Menozzi et al. 1978; Rosser et al. 2000; Sokal et al. 1991). These frequency gradients have been interpreted as a signature of a demic diffusion model (Ammerman & Cavalli-Sforza 1984; Menozzi et al. 1978), but some authors have argued they could have been created by the arrival of the first hunter-gatherers in Europe (Barbujani & Bertorelle 2001; Richards et al. 1996), although this hypothesis has never been formally tested. These two causes of gradient formation are actually difficult to distinguish since the first Paleolithic populations colonized Europe 40,000 years ago using approximately the same path as the Neolithization process 10,000 years ago (Bocquet-Appel & Demars 2000). The pattern of mtDNA diversity in European populations has been shown to be compatible with an old Paleolithic spatial expansion (Excoffier 2004; Ray et al. 2003), while evidence is contradictory for Y chromosome data. On one hand, clines of allele frequencies have been observed for several Y chromosome SNPs (Rosser et al. 2000) and a gradient of decreasing Neolithic contribution to the current gene pool has been inferred from the Near-East to the West by the analysis of 22 Y chromosome SNPs (Chikhi et al. 2002; Semino et al. 2000), in keeping with the hypothesis of a movement of Neolithic populations from the Near East and a progressive dilution of their gene poll by the incorporation of some Paleolithic lineages. On the other hand, the mismatch distributions of European populations inferred from the analysis of 22 Y chromosome SNPs do not show the typical signature of a demographic or spatial expansion (Pereira et al. 2001), which could be due to a small effective population size of males compared to females (potentially due to polygyny, Dupanloup et al. 2003), or to reduced male migration rates.

In order to assess the pattern of SNP diversity expected after the Neolithic expansion for various degrees of interactions with Paleolithic populations, we have carried out simulations of a range expansion in a spatially explicit model of Europe and the Near-East. These simulations were used to investigate three particular aspects of SNP diversity that have produced contradictory results discussed above: the existence of gradient of allele frequencies along a European South-East to North-West axis, the proportion of the European gene pool being of Paleolithic origin, and the

mismatch distribution within populations. Because an ascertainment bias in favor of SNP showing a relatively frequent minor allele is common (i.e. Casalotti et al. 1999) and leads to biased estimates of the past demography of a population (e.g. Wakeley et al. 2001), we have also examined its impact on patterns of molecular diversity.

Material and Methods

As reported previously (Excoffier 2004; Ray et al. 2003), realistic simulations of genetic diversity were carried out by first generating the forward demographic history (densities and migration rates between adjacent demes) of the populations. These demographic information are stored in a database, which is then used to generate the genealogies of samples of genes drawn in a predefined set of demes using a backward coalescent approach (e.g. Hudson 1990; Nordborg 2001).

Demographic simulations

While our approach is inspired by previous simulation studies on allele frequencies (e.g. Barbujani et al. 1995; Rendine et al. 1986), we have specifically modeled the occurrence of SNP mutations, and we have added some level of realisms, such as the spatial dynamics of Paleolithic populations and an explicit competition for local resources between Paleolithic and Neolithic populations. The spatial expansion of modern humans (*Homo sapiens sapiens*) in Europe, as well as the Neolithic transition were simulated using a modified version of the SPLATCHE program (Currat et al. 2004) as follows.

Digital model: A digital model of Europe and the Near-East has been created by dividing the continental surface in demes arranged on a grid. Each deme covers a surface of 50 by 50 km (or 2,500 km2), so that the modelled area has slightly more than 7,000 demes.

Range expansions: The colonization of Europe is assumed to have occurred in two phases. The first Paleolithic wave is assumed to have started some 1,600 generations ago (40,000 years ago with a generation time of 25 years) from the Near East (point P on Figure 1). This point has been chosen arbitrarily, as the source of modern humans having colonized Europe is not known exactly (Djindjian et al. 1999; Kozlowski & Otte 2000). A second colonization wave is assumed to have started from Anatolia (point N on Figure 1, Lev-Yadun et al. 2000) some 400 generations ago (corresponding to 10,000 years ago). At this time, the individuals occupying this deme are assumed to become farmers, and are moved in a new layer of 7,000 demes denoted as farmer or *F* demes, and superimposed on the layer of hunter-gatherers or *HG* layer.

Demographic regulation: The demography of more than 14,000 demes representing Europe (half in *HG* and half in *F* layers) is thus simulated during 1,600 generations, according to a model initially developed to describe the interactions between Neanderthals and modern humans (Currat & Excoffier 2004). In brief, density is logistically regulated within each deme (either belonging to the *F* or *HG* layer, and noted *i* below), with intrinsic rate of growth r_i and carrying capacity K_i. The local growth is also regulated by a density-dependent competition exerted by the population from the other layer competing for local resources, according to a modified version of the Lotka-Volterra model (see Currat & Excoffier 2004, for details). Each generation, a proportion *m* of individuals from

any given deme migrates to the neighboring demes from the same layer. At equilibrium, the local density N_i is equal to K_i, and the number of migrants exchanged between deme is thus equal to $K_i m$, which will be called $N_i m$ for coherence with previous work (e.g. Ray et al. 2003). *HG* contribution to the current genetic pool is simulated by a movement from the *HG* layer towards the *F* layer. This movement can be due to two processes: 1) adoption of Neolithic techniques by HG, a process also called acculturation (Ammerman & Cavalli-Sforza 1984) or 2) matings between Paleolithic and Neolithic individuals. The children resulting from these two processes are assumed to belong to the *F* layer and have thus an *HG* ancestor at the former generation. In the case of interbreeding, the amount of gene flow (*A*) between the two layers depends on the density of the individuals in layer *F* and *HG* in a given deme as $A = \gamma\left(2N_F N_{HG}\right) / \left(N_F + N_{HG}\right)^2$ deme, where γ controls the fecundity of the matings between individuals of the two layers. As discussed below, a pure *DD* model assumes that there was no genetic interaction between hunter-gatherers and farmers and therefore that γ=0. In that case, previous hunter-gatherers go extinct only due to their competition with Neolithic people. Less extreme demic diffusion models have been implemented, corresponding to different values of 0<γ<1, as reported in Table 1. The value of γ = 1 corresponds to the maximum amount of gene flow that can be simulated in our model and means that *HG* individuals reproduce indistinctly with *HG* or *F* individuals. It corresponds to the movement of 20 *HG* lineages per deme on average over the whole Europe. As a limiting case, a pure cultural transition was also simulated for which the *F* layer does not exist and where K_{HG} was simply multiplied by 20 within each deme. This demographic increase began at time -400 generations and was applied gradually from the Neolithic source deme at a speed corresponding to the scenario with γ = 0.

Parameter calibration: We gauged the parameters of our model from available paleo-demographic information. The carrying capacity of male or female hunter-gatherers (K_{HG}) before the Neolithic was set to 40, corresponding to a density of 0.064 individuals per km^2 (Alroy 2001; Steele et al. 1998). As it is largely accepted that the Neolithic transition coincides with the beginning of a significant increase in the population size (Bocquet-Appel & Dubouloz 2003; Cavalli-Sforza & Feldman 2003; Hassan 1979; Landers 1992), we have set K_F to 800, a value 20 times larger than K_{HG}. As *K* represents here the effective number of gender-specific genes (mitochondrial or Y chromosome), the total density simulated for the 5,500 demes constituting Europe is about 880,000 *HG* and 15 million farmers which are in broad agreements with the estimated number of people living in the Paleolithic and the Neolithic in Europe, respectively (Biraben 2003). Note also that K_F values larger than 800 do not affect the results substantially (results not shown). While it has been estimated that 500 generations were necessary for *HG* to colonize Europe (Bocquet-Appel & Demars 2000), the Neolithic transition was much more rapid, and took roughly between 4,000 and 8,000 years (Mazurié de Keroualin 2003; Price 2000), corresponding to 160-320 generations with a generation time of 25 years. These colonization times were used to calibrate the growth (*r*) and migration (*m*) rates. Values of r_{HG} = 0.4, r_F = 0.8, and *m* = 0.25 give colonization times in good agreement with figures mentioned above (see Table 1). Note that a growth rate of 80% per generation is very high but is within the upper range of rates considered as plausible for the human species (Ammerman & Cavalli-Sforza 1984; Pennington 2001; Young & Bettinger 1995). A migration rate of *m*=0.25 imply

the exchange of 10 males or 10 females between neighboring *HG* demes per generation and 200 individuals between *F* demes, two values in broad agreement with those estimated from mtDNA diversity in *HG* and post-Neolithic populations (Excoffier 2004).

Genetic simulations

We have simulated the diversity of samples of 40 genes in 20 demes located along an axis between the Near-East to Ireland (see Figure 1a.). For each reconstructed genealogy, the local Neolithic contribution to the current gene pool is measured as the proportion of sampled lineages whose ancestors belong to the source deme *F* at generation -400. In order to be able to compare our simulations with the Y chromosome data published for the European populations by Semino *et al* (2000) and in derived analyses (Dupanloup et al. 2003; Pereira et al. 2001), we have simulated 22 linked SNPs assumed to be on the Y chromosome. In order to detect allele frequency clines (*AFC*s), the frequency of the SNP is measured in each of the 20 simulated samples, and a linear regression is carried out over geographical distance between samples. If the regression coefficient is statistically significant at the 5% level, we consider this SNP as showing an *AFC*. The determination coefficient R^2 of the regression is also calculated for every statistically significant cline. In order to simulate different amounts of ascertainment bias, we have conducted separate analyses on SNPs with overall minor allele frequency among the 20 samples of at least 5% or at least 10%. The molecular diversity of a mtDNA sequence of 300bp was also simulated for the same samples, assuming a mutation rate of 0.00125 per generation for the whole sequence (≈ 33 % of divergence par million years, (Heyer et al. 2001; Soodyall et al. 1997). The genetic variability of the samples was analyzed using the program ARLEQUIN (Schneider et al. 2000).

Results

Distinction between cultural (CD) and demic (DD) diffusion models

The molecular signature obtained under various scenarios depends on the spatio-temporal dynamics of the sampled lineages. Under a pure *DD* model (without genetic exchange between Neolithic and Paleolithic populations, γ=0), and going backward in time, the ancestors of the sampled lineages first coalesce or disperse in the *F* layer (Figure 1a). Then, they are brought back to the place of origin of the Neolithic expansion by the shrinking Neolithization wave (Figure 1b-c). Some of them pass through the spatial and demographic bottleneck constituted by the Neolithic source. The lineages that did not coalesce during this bottleneck can disperse again in the *HG* layer (Figure 1d). Finally, the lineages are brought back towards the place of origin of the Paleolithic expansion (Figure 1e-f). This dynamics results in three main periods of coalescent events: the "scattering” phase (sensu Wakeley 1999, *S1* in Figure 2), followed by two "contraction" phases (corresponding to range expansions when going forward in time), that respectively take place during the Neolithic (*C1*) and the Paleolithic (*C2*) migration waves. As illustrated on Figure 2, the relative proportion of coalescent events taking place during the two “contraction” phases *C1* and *C2* are quite different under the pure *DD* model (γ=0) and with high Paleolithic input (γ=1). The number of coalescent events in the scattering phase *S1* only depends on the parameter $N_F m$, as shown

previously (Ray et al. 2003), and it does not allow one to distinguish between the two models. It thus appears that the period *C1* is critical to distinguish between models. Under a pure *DD* model, almost all coalescent events (98%) occur before the lineages reach the initial Neolithic deme (Figure 2). Contrastingly, only about half (49%) of the coalescent events occur after the onset of the Neolithic transition when γ=1. Under this latter case, less than 10% of the coalescent events occur within the Neolithization wave and 20% within the Paleolithic contraction wave *C2* (Figure 2). The remaining 70% occur in the *HG* layer during or before Neolithic times, after the passage of the Neolithic wave because the lineages evolve in demes with low densities. Note that the number of coalescent events occurring within the Neolithization front depends on γ, the amount of gene flow between the two layers, so that smaller γ values translate into larger numbers of coalescent events. The number of migrants exchanged between demes from the *HG* layer ($N_{HG}m$) does not affect the genetic pattern (results not shown), and low $N_{HG}m$ values only slightly increase the number of coalescent events that occurs within the *HG* population. The influence of r_{HG} on the coalescent tree is negligible.

Importance of the migration front

Our simulations underline the role of the range expansion processes for generating *AFC*s. The colonization process corresponds to a succession of founder effects occurring at the wave front (Austerlitz et al. 2000). In a coalescent perspective, the lineages that are spread over a wide area are gathered and concentrated by the contracting wave front, and have thus an increased probability to coalesce during the contraction of the occupied territory. Our simulations reveal that *AFC*s are extremely rare for randomly chosen SNPs, but that they become very frequent in case of an ascertainment bias consisting in selecting SNPs with minor allele frequencies larger than 5% (Table 1). Since gene genealogies resulting from a range expansion have usually long terminal branches (Excoffier 2004; Ray et al. 2003), SNP mutations will most of the time occur on these terminal branches and will consist in singletons when the number of migrants exchanged between neighboring demes is large, or could reach low frequencies but be geographically restricted when migration is lower. Therefore, randomly chosen SNPs will generally not show clinal patterns since they will be spread over a small region. With ascertainment bias, the fraction of SNPs showing *AFC*s increases dramatically, and can even be observed in about 50% of the loci (Table 1). Interestingly, the *AFC*s occur at about the same frequency, independently of the amount of incorporation of Paleolithic lineages into the *F* layer (Table 1), and thus at similar frequencies under a pure *DD* or a pure acculturation model. It implies that *AFC*s cannot be considered as indicative of a range expansion of Neolithic farmers, since they could have been created equally well during the first expansion of modern humans into Europe.

The Neolithisation front is also important because it is the region where *HG* and *F* demes coexist, and consequently where genetic exchanges occur between the two layers. Therefore, the probability for a lineage to be of *HG* ancestry increases with the time spent within the Neolithization front during the contraction periods *C1* or *C2*. The proportion of lineages whose ancestors trace back to the *F* layer diminishes rapidly with increasing distance from the Neolithic source (Figure 3). Obviously, when γ increases the total proportion of Neolithic lineages decreases, and these lineages

are restricted to the area of the origin of the Neolithic (Figure 3). Even when γ=1, there is still 1% of "Neolithic lineages" in the Anatolian sample close to the source of the Neolithic. Note that, under the simulated conditions, a Neolithic cline is observed at the continental level only when γ is smaller than 0.15 (corresponding to about 3 *HG* incorporated per deme on average). It is also important to note that even for values of γ as low as 0.05 (1 *HG* incorporated per deme during the whole cohabitation period) the majority of the current European gene pool is of Paleolithic ancestry (Table 1, Figure 3). This results is virtually not affected by the size and the spread of the Neolithic source, for instance when it consists of a subdivided population of 25 demes (Currat 2004).

Molecular diversity within demes

The patterns of molecular diversity can be obtained by adding mutations on top of coalescent trees, whose structures are readily perceived in Figure 2 for γ = equal to 0 and 1. Under a pure *DD* model (γ =0), a large proportion of mismatch distributions are multimodal, have a large variance, and present an important proportion of identical pairs of sequences (Figure 4a-b). The homozygosity (class 0 in mismatch distributions) increases with the distance between the sampling area and the Neolithic source, because the number of coalescent events occurring during the *C1* phase will also increase. When γ increases, the difference between samples located close or far from the Neolithic source disappears, and the proportion of unimodal mismatch distributions quickly increases (~50% with γ = 0.05 and ~90% with γ = 0.15) and is close to 95% when γ > 0.5 (Figure4c-d). This increase in the number of unimodal mismatch is faster for populations which are furthest away from the Neolithic source since it is also those integrating the most Paleolithic genes. The mismatch distributions simulated for 22 SNPs when γ = 0 are often bimodal, whereas they are almost always unimodal when γ = 1 (Figure 5a-b). As soon as ascertainment bias is introduced, the realized mismatch distributions become multimodal under all simulated scenarios (Figure 5c-d), even though the average distributions are relatively flat.

Discussion

Simulating a realistic Neolithic range expansion

The degree of realism of our simulations of the colonization of Europe by Homo sapiens sapiens followed by a second Neolithic range expansion is difficult to judge, as the true history of the European population has certainly been even more complex (Mazurié de Keroualin 2003). However, these simulations are more realistic than those done previously (Barbujani et al. 1995; Rendine et al. 1986), and fit the known duration of the Neolithic transition process as well as the duration of the Mesolithic period in several places. Since simulated cohabitation times between *HG* and *F* demes vary between 5.6 and 7.7 generation (150 to 200 years) (Table 1), they are thus close to documented cases where the two types of economies coexisted over larger areas, like 300-700 years in the North of the Alps and the Jura: (Gallay 1994), 800 years in Cantabria and 400 years in Portugal (Arias 1999), or 200 years in Franche-Compté (Jeunesse 1998).

Our simulations were performed in a homogeneous environment with γ identical in every deme, regardless of its location. While this assumption may seem unrealistic at a regional scale, it is quite reasonable at a continental scale since the speed of *HG* colonization and that of the Neolithic transition can be regarded as quite regular at this level (Ammerman & Cavalli-Sforza 1984; Bocquet-Appel & Demars 2000). It would be interesting to test in future studies the influence of some heterogeneity of the migration wave, and to incorporate, with much additional work and computer power, more realism in the simulation, such as an heterogeneous environment subject to temporal fluctuations (Adams & Faure 1997), spatial heterogeneity in γ inferred from archaeological information (Lahr et al. 2000), maritime migrations along the Mediterranean coasts (Zilhao 2001), or contractions/re-expansion during ice ages and long distance dispersal. It however appears necessary to understand the genetic signature expected under a relatively simple demographic scenario, before considering more complex ones.

Allele Frequency clines and influence of ascertainment bias

Allele frequency clines (*AFC*) can be generated by a succession of founders effects along the axis of diffusion of an expansion wave (Austerlitz et al. 2000; Barbujani et al. 1995; Fix 1997). However, our results show that such clines can only be generated for alleles that are selected to be relatively frequent over the whole range of the studied area. It therefore suggests that these clines will only be observed for alleles that are older than - or that have occurred in the initial phase of - the expansion (possibly at the front of the wave of advance, Edmonds et al. 2004). In that sense, an ascertainment bias in favor of SNPs with frequent minor alleles will show frequency clines in about 50% of the cases after an expansion (Table 1), whereas no or a non-significant number of clines will be observed without ascertainment bias. This difference can perhaps explain the fact that *AFC*s have been commonly observed for classical markers (Menozzi et al. 1978; Sokal et al. 1991), STR and SNPs (i.e. Chikhi et al. 1998; Rosser et al. 2000) in Europe, but not for mtDNA when unascertained complete sequence data are used (Richards et al. 1996; Richards et al. 1998). Note that when ascertainment is artificially exerted on mtDNA sequence, for instance by defining haplogroups on the basis of old mutations defining mtDNA lineages, a geographic structure and gradient of haplogroup frequencies begins to be observed (Richards et al. 2002).

Our simulations suggest that *AFC* from the Middle East to North-western Europe can be generated equally well by the Neolithic expansion process that occurred 8,000-3,000 BC or by the expansion of the first modern human in Europe ~45,000-30,000 BP. It is important to recognize that *AFC*s are not generated by the different amounts of Paleolithic lineages in the current demes along the expansion path (Figure 3), since clines are present even in total absence of such lineages, as in the case of a pure *DD* model (γ = 0). In fact, the occurrence of these *AFC*s is relatively independent of the contribution of Paleolithic lineages into the current gene pool of Europeans (Table 1). The expected frequency of *AFC*s under a pure cultural diffusion (when the *F* layer does not exist, i.e. Table 1, last line) is even larger than under pure *DD* model (γ = 0), due to the fact that founder effects are stronger in small populations. Since the presence of *AFC*s is thus independent of the proportion of Neolithic lineages in the population, they cannot be invoked as a pure support to the

DD theory (Barbujani & Bertorelle 2001; Barbujani et al. 1995), and only the dating of the *AFC*s would perhaps allow the support of one model rather than another.

Paleolithic contribution to the European genetic pool

The nature of the founders of a population is important to determine its final genetic composition (Heyer 1995; Heyer & Tremblay 1995; Milinkovitch et al. 2004), because the majority of individuals present at equilibrium are descendants from the first colonists (Currat & Excoffier 2004; Edmonds et al. 2004). Our simulations show that a very small initial Paleolithic contribution in each deme (0.125% on average) is enough to lead to a situation where most of the current gene pool can be traced to the Paleolithic (Table 1). The proportion of European who are descendant from the first farmers from the Levant decreases very quickly with distance from the Neolithic source, as the lineages of Neolithic origin are rapidly diluted along the axis of colonization (Figure 3). Under our simulation conditions, an average local Paleolithic contribution larger than 0.375%, will indeed be enough to prevent Neolithic lineages to diffuse over the whole Europe.

These results imply that, under our model of a progressive range expansion of Neolithic farmers with possible genetic exchange and competition with local Paleolithic hunter-gatherers, it is very unlikely that the Paleolithic contribution be globally smaller than 50%. If that was the case (e.g. Chikhi et al. 2002, < 30%), it would imply that Neolithic would have had virtually no genetic contact with local populations, like under a pure *DD* model. Global surveys of mtDNA molecular diversity (Richards et al. 1996; Richards et al. 2000), and the simulations of mtDNA mismatch distributions argue against such a low contribution of Paleolithic populations to the modern gene pool. Indeed, examination of Figure 4, reveals that in absence of exchange with hunter-gatherers, mismatch distributions should often be multimodal, and have a mode closer to zero in populations sampled far from the Neolithic source. On the contrary, most European mismatch distributions are smooth and unimodal (Excoffier & Schneider 1999), and the mode of mismatch distributions is quite homogeneous across Europe (Excoffier 2004), as expected when the contribution of Paleolithic lineages becomes important. Moreover, previous dating of demographic expansion for European populations pointed towards 40,000 years ago or more (Comas et al. 1996; Excoffier & Schneider 1999), in keeping with a Paleolithic expansion.

Influence of ascertainment bias on SNP diversity

Ascertainment bias has also a drastic effect on the shape of mismatch distributions inferred from linked SNPs, as they become highly multimodal for relatively large amounts of acertainment bias (minor allele frequency > 10%). Therefore, this kind of ascertainment bias can erase a signature of demographic or range expansion. It is interesting to note that it is precisely the conclusion that was drawn from the analysis of 22 linked Y chromosome SNPs showing bimodal mismatch distribution (Pereira et al. 2001),where the absence of expansion signal was attributed to a smaller male than female effective size (Dupanloup et al. 2003). Note however, that bimodal mismatch distributions can also be obtained under a pure *DD* model (Figure 5a), but this model was shown above to be unlikely from the analysis of mtDNA. It follows that observed differences between the mismatch

distributions obtained from mtDNA sequences and from Y chromosome SNPs can be explained by the mere selection of frequent Y chromosome SNPs, which is also supported by the observation of allele frequency clines for these markers and not for mtDNA sequences.

Acknowledgements

Thanks to Nicolas Ray and Pierre Berthier for their programming and computing assistance. We are also grateful to Montgomery Slatkin and Estella Poloni for stimulating discussion on the subject and to Grant Hamilton for his careful reading of the manuscript. This work was supported by a Swiss NSF grant No 3100A0-100800 to LE.

Table & Figures

Table 1. Statistics computed after the simulation of various amount of interactions between Paleolithic and Neolithic populations.

Paleolithic contribution		*Demographic variables*			*Neolithic Contribution* [f]	*Allele Frequency Clines* [g]					
						No bias		*Bias (5%)*		*Bias (10%)*	
γ [a]	L[b]	*HG col.* [b]	*F col.* [d]	*Cohab* [e]		*Freq.*	*R^2*	*Freq*	*R^2*	*Freq.*	*R^2*
0.00	0	470	260	7.7	1.00 *0.00*	0.03	*0.50*	0.57	*0.60*	0.56	*0.62*
0.05	1	470	260	7.7	0.48 *0.13*	0.03	*0.47*	0.48	*0.54*	0.45	*0.58*
0.10	2	470	255	7.6	0.30 *0.10*	0.03	*0.45*	0.50	*0.56*	0.51	*0.63*
0.15	3	470	250	7.4	0.12 *0.04*	0.04	*0.42*	0.51	*0.58*	0.78	*0.70*
0.25	5	470	245	7.3	0.07 *0.02*	0.03	*0.42*	0.66	*0.59*	0.86	*0.71*
0.50	10	470	240	7.0	0.03 *0.01*	0.02	*0.43*	0.71	*0.58*	0.82	*0.68*
0.75	15	470	230	6.7	0.01 *0.00*	0.02	*0.40*	0.70	*0.58*	0.82	*0.67*
1.00	20	470	220	5.6	0.00 *0.00*	0.02	*0.40*	0.68	*0.59*	0.80	*0.63*
-	-*	470	260°	-	0.00	0.02	*0.40*	0.68	*0.58*	0.78	*0.66*

[a] γ : rate of gene flow between *HG* and *F* demes. Minimum = 0 (no gene flow) and maximum = 1.0.

[b] L: Average number of Paleolithic lineages incorporated per deme.

[c] and [d]: Colonization time of Europe by Paleolithic and Neolithic range expansions, respectively.

[e] Mean cohabitation time (in generation) between *HG* and *F* within a deme.

[f] Average "Neolithic" contribution to the current European genetic pool (see text) over 10,000 simulations, standard deviation are shown in italic.

[g] Freq.: proportion of simulation (over 10,000) that show a significant *AFC* at the 5% significance level, R^2 = average determination coefficient for the significant *AFCs*. * Only one population is simulated. ° Time for cultural diffusion over whole Europe

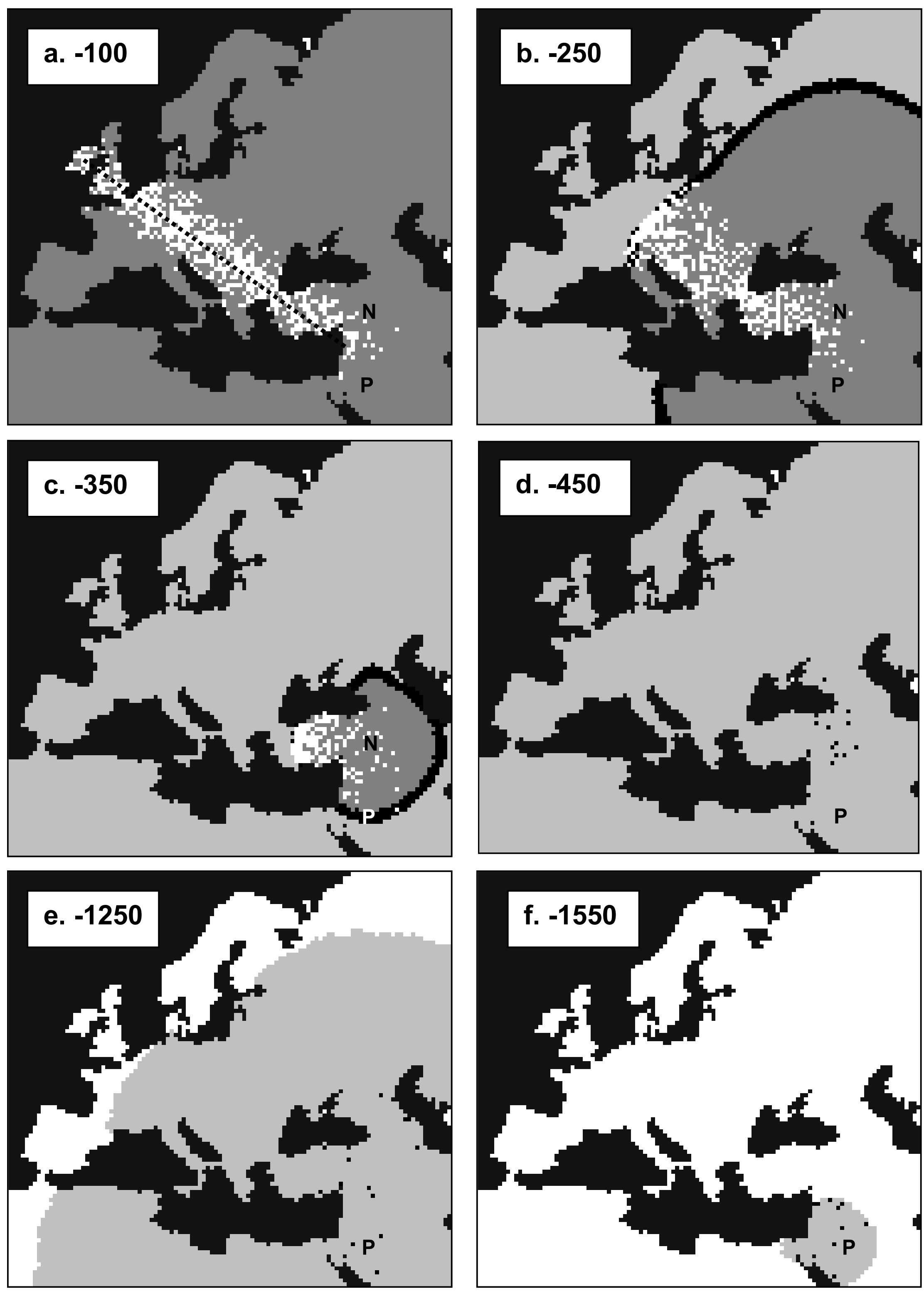


Figure 1: Spatial and temporal dynamics of the location of ancestral lineages under a double Neolithic and Paleolithic range expansion from the Near-East. The six panes a) to f) show the location of ancestral lineages and the area occupied by Neolithic (layer *F*,in dark gray) and Paleolithic (layer *HG* in light gray) demes at six different periods before present under a pure *DD* model (γ = 0). P = origin of the Paleolithic expansion and F = origin of the Neolithic expansion. Dashed lines = the axe along which 20 demes are samples for 40 genes. Black spots on the light gray zone represent *HG* lineages and white spots on the dark gray zone represent *F* lineages. The black band at the front of the Neolithic expansion represents the Mesolithic zone where Neolithic and Paleolithic populations coexist.

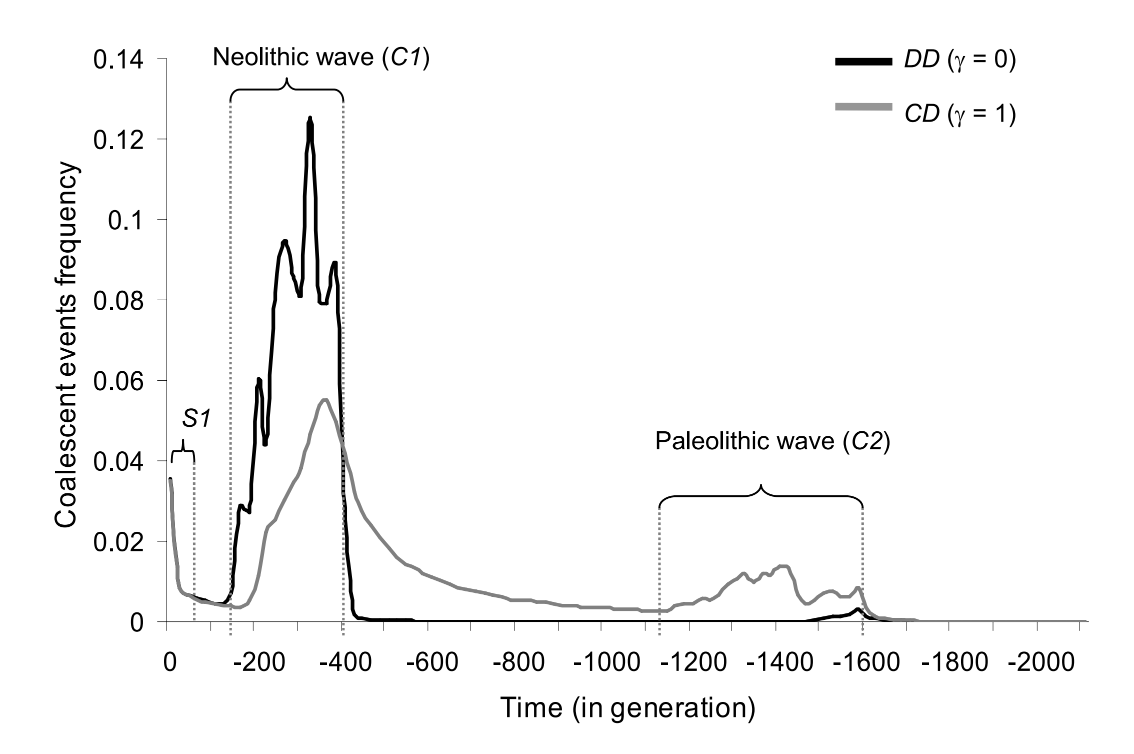


Figure 2: Temporal distribution of the coalescent events under the pure *DD* model (γ=0, when there is no genetic interaction between hunter-gatherers and farmers, in black) and when γ=1 (the maximum amount of gene flow allowed, in grey). *S1* correspond to the "scattering" phase (see text), *C1* and *C2* to the "contraction" phases occurring during the Neolithic and the Paleolithic expansions, respectively. The small variations in the distributions are due to spatial bottlenecks (Currat 2004).

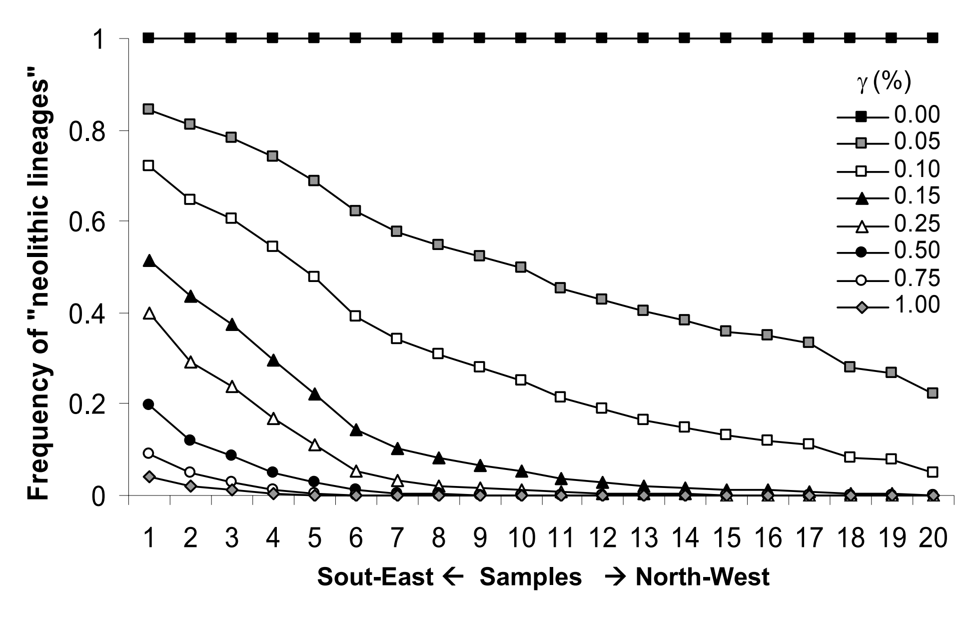


Figure 3: Proportion of "Neolithic lineages" in every sample from the Near East (1) to the North West (20) of Europe, for rates of gene flow between *HG* and *F* (γ).

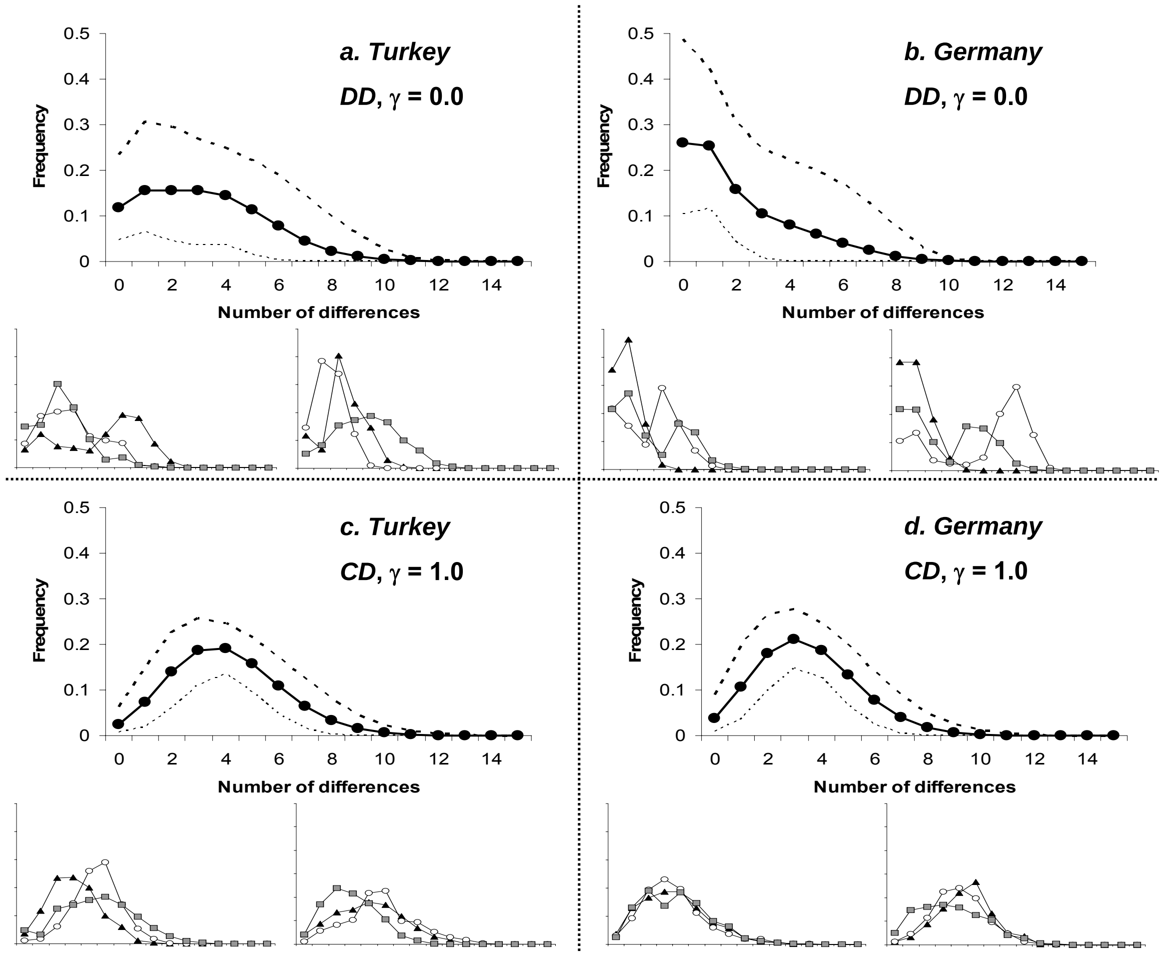


Figure 4: Expected mismatch distributions obtained from 10,000 genetic simulations of 300 bp DNA sequences for samples located in Turkey and in Germany, without or with maximum genetic flow between *HG* and *F*. Dashed lines correspond to the limits of a 90% confidence interval for the mismatch distribution. Small graphs show 6 independent replicates of each case studied here. $N_F m$ = 200.

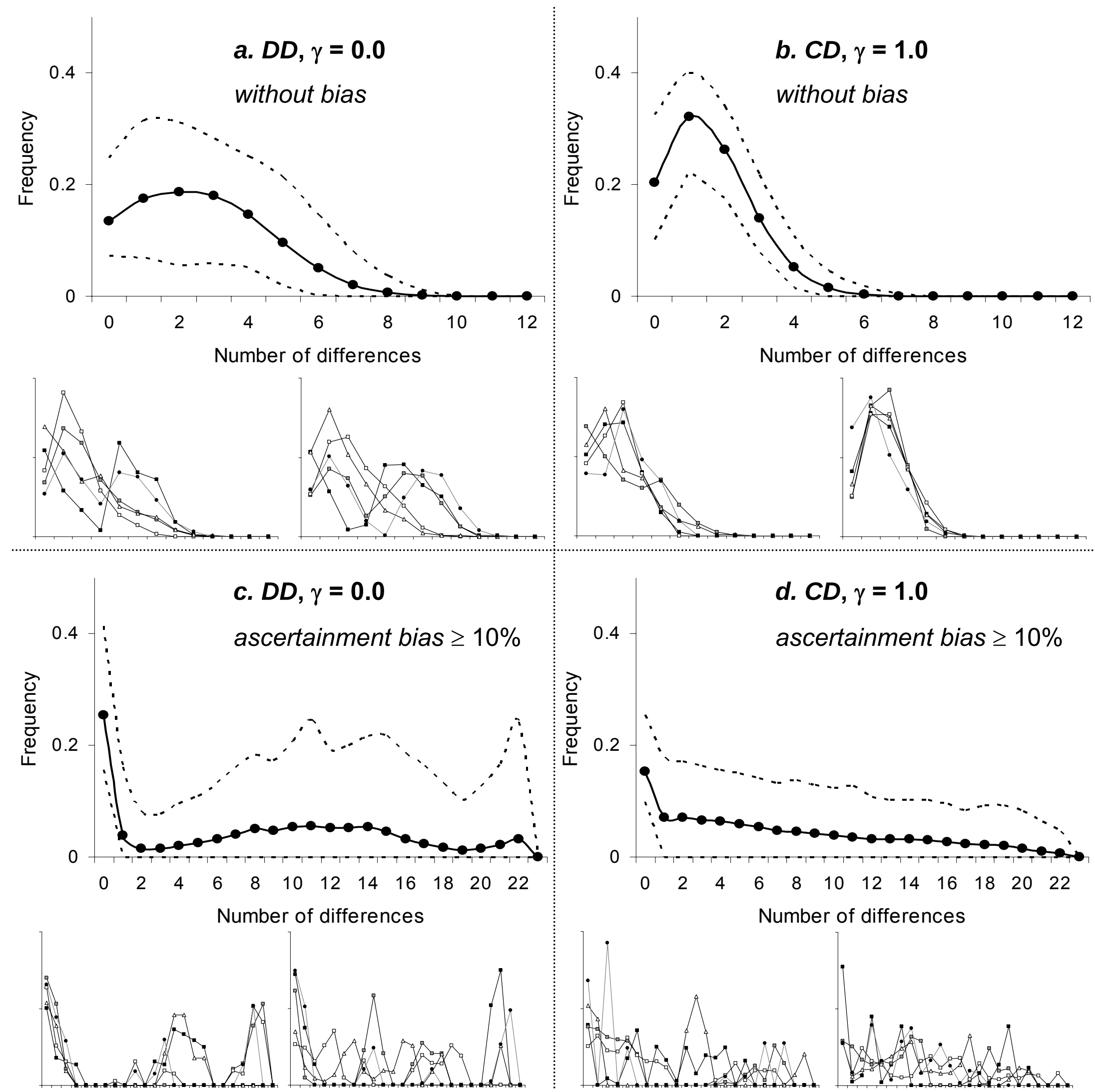


Figure 5: Expected mismatch distributions obtained from 10,000 genetic simulations of 22 linked SNPs for samples located in Germany without or with maximum genetic flow between *HG* and *F*, with and without ascertainment bias. Dashed lines correspond to the limits of a 90% confidence interval for the mismatch distribution. Small graphs show 10 independent replicates of each case studied here. Ascertainment bias was modeled by selecting SNPS with a minor allele frequency exceeding 10% along the transect shown on Figure 1.

Additional Figures

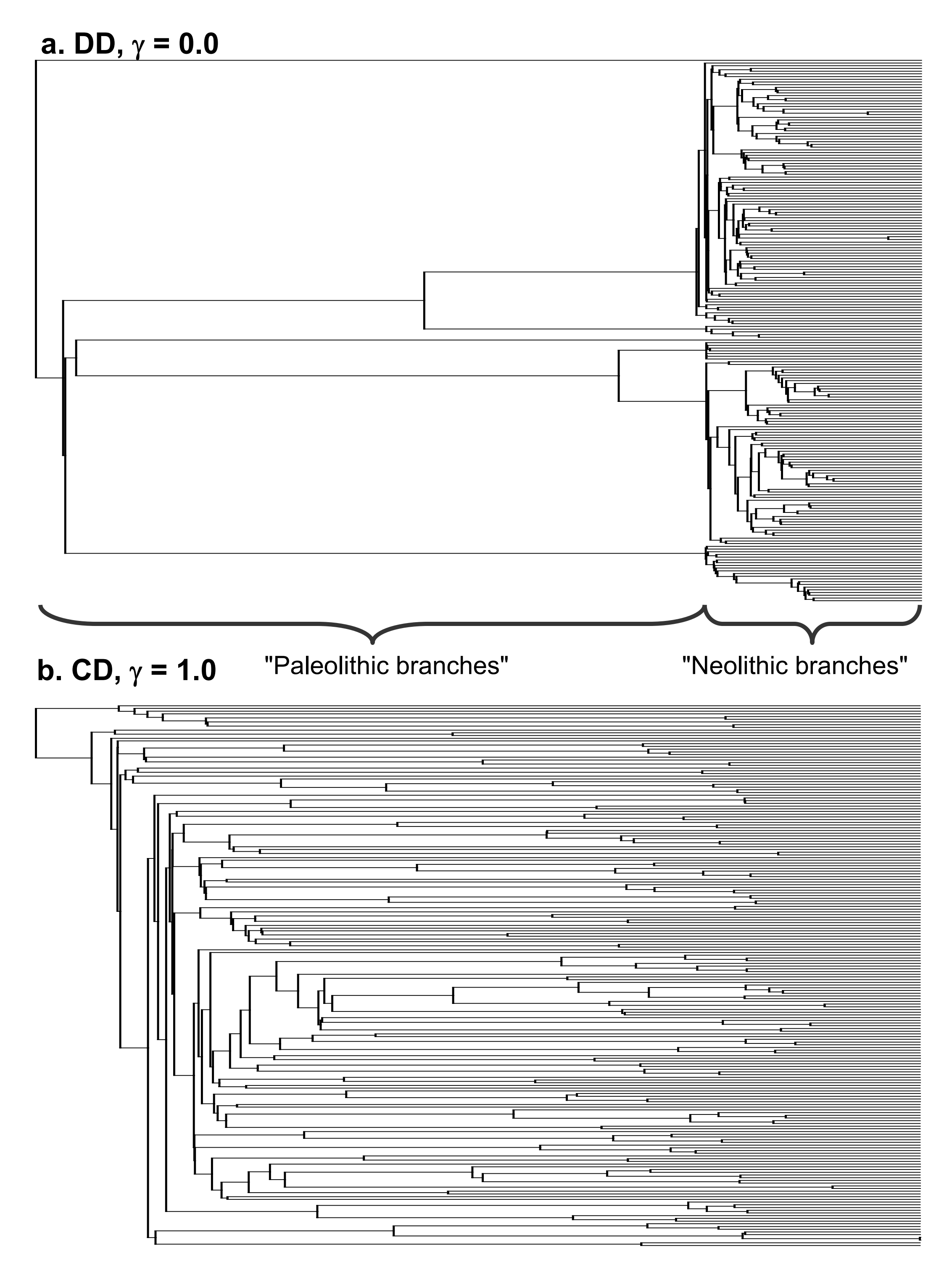


Figure S6: Typic genealogical trees simulated under *DD* (a) and *CD* (b).

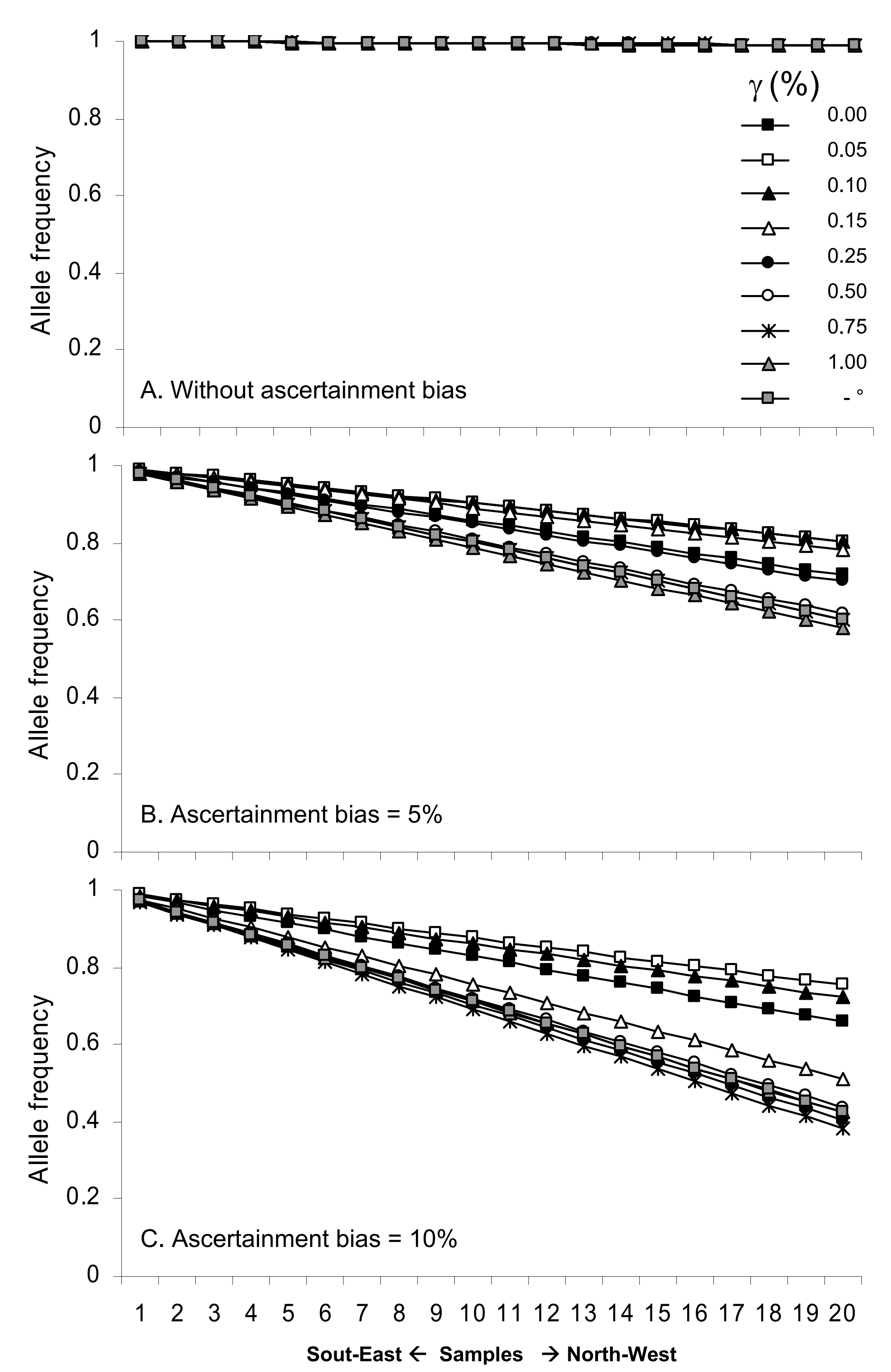


Figure S7: Mean linear regression of allele frequencies on geographic distances, over 10,000 simulations, with rates of gene flow between *HG* and *F*. A) without ascertainment bias; B) with ascertainment bias equal to 5%. and B: 10%. ° Only one population is simulated.

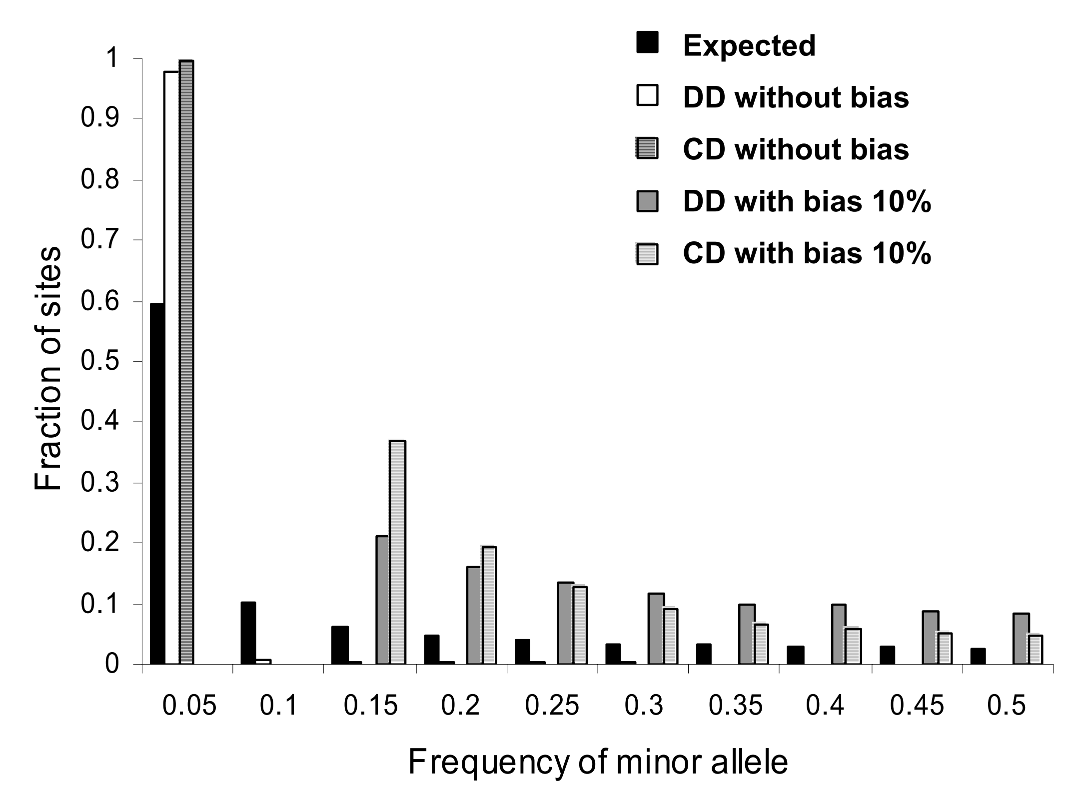


Figure S8 : Allele frequency spectrum Harpending *et al.* 1998 under DD (γ = 0.0) and *CD* (γ = 1.0), with or without ascertainment bias. Black bands are the expected values under neutrality and constant population size Tajima 1989a.

References

Adams, J. & Faure, H. 1997 Review and atlas of palaeovegetation: preliminary land ecosystem maps of the world since Last Glacial Maximum: Oak Bridge National Laboratory.

Alroy, J. 2001 A multispecies overkill simulation of the end-Pleistocene megafaunal mass extinction. Science 292, 1893-1896.

Ammerman, A. & Cavalli-Sforza, L. L. 1984 The Neolithic transition and the genetics of populations in Europe. Princeton, New Jersey: Princeton University Press.

Arias, P. 1999 The origins of the Neolithic along the Atlantic coast of continental Europe. Journal of World Prehistory 13, 403-464.

Austerlitz, F., Mariette, S., Machon, N., Gouyon, P. H. & Godelle, B. 2000 Effects of colonization processes on genetic diversity: differences between annual plants and tree species. Genetics 154, 1309-21.

Barbujani, G. & Bertorelle, G. 2001 Genetics and the population history of Europe. Proc Natl Acad Sci U S A 98, 22-5.

Barbujani, G. & Dupanloup, I. 2002 DNA Variation in Europe: estimating the demographic impact of Neolithic dispersals. In Examining the farming/language dispersal hypothesis (ed. P. Bellwood & C. Renfrew), pp. 421-431. Cambridge: McDonald Institute Monographs.

Barbujani, G. & Pilastro, A. 1993 Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. Proceedings of the National Academy of Science 90, 4670-3.

Barbujani, G., Sokal, R. R. & Oden, N. L. 1995 Indo-European origins: a computer-simulation test of five hypotheses. Am J Phys Anthropol 96, 109-32.

Biraben, J.-N. 2003 L'évolution du nombre des hommes. Population et Sociétés 394, 1-4.

Bocquet-Appel, J.-P. & Demars, P. Y. 2000 Neanderthal contraction and modern human colonization of Europe. Antiquity 74, 544-552.

Bocquet-Appel, J.-P. & Dubouloz, J. 2003 Traces paléoanthropologiques et archéologiques d'une transition démographique néolithique en Europe. Bulletin de la société préhistorique française 100, 699-714.

Casalotti, R., Simoni, L., Belledi, M. & Barbujani, G. 1999 Y-chromosome polymorphisms and the origins of the European gene pool. Proc R Soc Lond B Biol Sci 266, 1959-1965.

Cavalli-Sforza, L. L. & Feldman, M. W. 2003 The application of molecular genetic approaches to the study of human evolution. Nat Genet 33 Suppl, 266-75.

Chikhi, L. 2002 Admixture and the demic diffusion model in Europe. In Examining the farming/language dispersal hypothesis (ed. P. Bellwood & C. Renfrew), pp. 435-447. Cambrigs: McDonald Institute Monographs.

Chikhi, L., Destro-Bisol, G., Bertorelle, G., Pascali, V. & Barbujani, G. 1998 Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. Proc Natl Acad Sci U S A 95, 9053-8.

Chikhi, L., Nichols, R. A., Barbujani, G. & Beaumont, M. A. 2002 Y genetic data support the Neolithic demic diffusion model. PNAS 99, 11008-11013.

Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A. & Bertranpetit, J. 1996 Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. Mol Biol Evol 13, 1067-77.

Currat, M. 2004 Effets des expansions des populations humaines en Europe sur leur diversité génétique. In Thesis, Département d'Anthropologie et Ecologie. Genève: Université de Genève.

Currat, M. & Excoffier, L. 2004 Model of range expansion of modern humans in Europe supports no admixture with Neanderthals. submitted.

Currat, M., Ray, N. & Excoffier, L. 2004 SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Molecular Ecology Notes 4, 139-142.

Djindjian, F., Koslowski, J. & Otte, M. 1999 Le Paléolithique supérieur en Europe. Paris: Armand Colin.

Dupanloup, I., Pereira, L., Bertorelle, G., Calafell, F., Prata, M. J., Amorim, A. & Barbujani, G. 2003 A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. J Mol Evol 57, 85-97.

Edmonds, C. A., Lillie, A. S. & Cavalli-Sforza, L. L. 2004 Mutations arising in the wave front of an expanding population. Proc Natl Acad Sci U S A 101, 975-9.

Excoffier, L. 2004 Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Mol Ecol 13, 853-64.

Excoffier, L. & Schneider, S. 1999 Why hunter-gatherer populations do not show sign of Pleistocene demographic expansions. Proceedings of the National Academy of Sciences USA 96, 10597-10602.

Fix, A. G. 1997 Gene frequency clines produced by kin-structured founder effects. Hum Biol 69, 663-73.

Gallay, A. 1994 A propos de travaux récents sur la Néolithisation de l'Europe de l'ouest. L'Anthropologie 98, 576-588.

Gronenborg, D. 1999 A variation on a basic theme: the transition to farming in southern central Europe. Journal of World Prehistory 13, 123-210.

Hassan, F. A. 1979 Demography and archaeology. Annual Review of Anthropology 8, 137-160.

Heyer, E. 1995 Mitochondrial and nuclear genetic contribution of female founders to a contemporary population in northeast Quebec. Am J Hum Genet 56, 1450-5.

Heyer, E. & Tremblay, M. 1995 Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. Am J Hum Genet 56, 970-8.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J. & Labuda, D. 2001 Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. Am J Hum Genet 69, 1113-26.
Hudson, R. R. 1990 Gene genealogies and the coalescent process. In Oxford Surveys in Evolutionary Biology (ed. D. J. Futuyma & J. D. Antonovics), pp. 1-44. New York: Oxford University Press.
Jeunesse, C. 1998 La néolithisation de l'Europe occidentale (VIIe-Ve millénaires av. J.-C.): nouvelles perspectives. In Les derniers chasseurs-cueilleurs du massif jurassien et de ses marges (ed. C. Cupillard & A. Richard). Lons-le-Saunier: Centre Jurassien du patrimoine.
Kozlowski, J. & Otte, M. 2000 The formation of the Aurignacian. Journal of Anthropological Research 56, 513-524.
Lahr, M. M., Foley, J. A. & Pinhasi, R. 2000 Expected regional patterns of mesolithic-neolithic human population admixture in Europe based on archaeological evidence. In Archaeogenetics:DNA and the population prehistory of Europe, vol. 1 (ed. C. Renfrew & K. Boyle), pp. 81-88. Cambridge: McDonald Institute for Archaeological Research, University of Cambridge.
Landers, J. 1992 Reconstructing ancient populations. In The Cambridge Encyclopedia of Human Evolution. (ed. S. Jones, R. Martin & D. Pilbeam), pp. 402-405. London: Cambrige University Press.
Lev-Yadun, S., Gopher, A. & Abbo, S. 2000 Archaeology. The cradle of agriculture. Science 288, 1602-3.
Mazurié de Keroualin, K. 2003 Genèse et diffusion de l'agriculture en Europe : agriculteurs, chasseurs, pasteurs. Paris: Errance.
Menozzi, P., Piazza, A. & Cavalli-Sforza, L. 1978 Synthetic maps of human gene frequencies in Europeans. Science 201, 786-92.
Milinkovitch, M. C., Monteyne, D., Gibbs, J. P., Fritts, T. H., Tapia, W., Snell, H. L., Tiedemann, R., Caccone, A. & Powell, J. R. 2004 Genetic analysis of a successful repatriation programme: giant Galápagos tortoises. Proc R Soc Lond B Biol Sci 271, 341-345.
Nordborg, M. 2001 Coalescent Theory. In Handbook of Statistical Genetics (ed. D. Balding, M. Bishop & C. Cannings), pp. 179-212. New York: John Wiley & Sons Ltd.
Pennington, R. 2001 Hunter-gatherer demography. In Hunter-gatherers: an interdisciplinary perspective (ed. C. Panter-Brick, R. H. Layton & P. Rowley-Conwy), pp. 170-204: Cambridge University Press.
Pereira, L., Dupanloup, I., Rosser, Z. H., Jobling, M. A. & Barbujani, G. 2001 Y-chromosome mismatch distributions in Europe. Mol Biol Evol 18, 1259-71.
Price, T. D. 2000 Europe's first farmers. Cambridge: Cambridge University Press.
Ray, N., Currat, M. & Excoffier, L. 2003 Intra-deme molecular diversity in spatially expanding populations. Mol Biol Evol 20, 76-86.
Rendine, S., Piazza, A. & Cavalli-Sforza, L. 1986 Simulation and separation by principal components of multiple demic expansions in Europe. Am. Nat. 128, 681-706.
Richards, M. 2003 The Neolithic invasion of Europe. Annu. Rev. Anthropol. 32, 135-162.
Richards, M., Corte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H. J. & Sykes, B. 1996 Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 59, 185-203.
Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Golge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C. & Stefanescu, G. 2000 Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67, 1251-76.
Richards, M., Macaulay, V., Torroni, A. & Bandelt, H. J. 2002 In search of geographical patterns in European mitochondrial DNA. Am J Hum Genet 71, 1168-74.
Richards, M. B., Macaulay, V. A., Bandelt, H. J. & Sykes, B. C. 1998 Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62 (Pt 3), 241-60.
Rosser, Z. H., Zerjal, T., Hurles, M. E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., Beckman, G., Beckman, L., Bertranpetit, J., Bosch, E., Bradley, D. G., Brede, G., Cooper, G., Corte-Real, H. B., de Knijff, P., Decorte, R., Dubrova, Y. E., Evgrafov, O., Gilissen, A., Glisic, S., Golge, M., Hill, E. W., Jeziorowska, A., Kalaydjieva, L., Kayser, M., Kivisild, T., Kravchenko, S. A., Krumina, A., Kucinskas, V., Lavinha, J., Livshits, L. A., Malaspina, P., Maria, S., McElreavey, K., Meitinger, T. A., Mikelsaar, A. V., Mitchell, R. J., Nafa, K., Nicholson, J., Norby, S., Pandya, A., Parik, J., Patsalis, P. C., Pereira, L., Peterlin, B., Pielberg, G., Prata, M. J., Previdere, C., Roewer, L., Rootsi, S., Rubinsztein, D. C., Saillard, J., Santos, F. R., Stefanescu, G., Sykes, B. C., Tolun, A., Villems, R., Tyler-Smith, C. & Jobling, M. A. 2000 Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67, 1526-43.
Schneider, S., Roessli, D. & Excoffier, L. 2000 Arlequin: a software for population genetics data analysis. User manual ver 2.000. Geneva: Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.
Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., Marcikiae, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A. S., Cavalli-Sforza, L. L. & Underhill, P. A. 2000 The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290, 1155-9.
Sokal, R. R., Oden, N. L. & Wilson, C. 1991 Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351, 143-5.

Soodyall, H., Jenkins, T., Mukherjee, A., du Toit, E., Roberts, D. F. & Stoneking, M. 1997 The founding mitochondrial DNA lineages of Tristan da Cunha Islanders. Am J Phys Anthropol 104, 157-66.

Steele, J., Adams, J. M. & Sluckin, T. 1998 Modeling Paleoindian dispersals. World Archeology 30, 286-305.

Wakeley, J. 1999 Nonequilibrium migration in human history. Genetics 153, 1863-71.

Wakeley, J., Nielsen, R., Liu-Cordero, S. N. & Ardlie, K. 2001 The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. Am J Hum Genet 69, 1332-47.

Young, D. A. & Bettinger, R. L. 1995 Simulating the global human expansion in the late pleistocene. Journal of Archaeological Science 22, 89-92.

Zilhao, J. 2001 Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe. Proc Natl Acad Sci U S A 98, 14180-5.

Zvelebil, M. 1986 Review of Ammerman & Cavalli-Sforza (1984). Journal of Archaeological Science 13, 93-95.

Zvelebil, M. & Zvelebil, K. V. 1988 Agricultural transition and Indo-European dispersals. Antiquity 62, 574-583.

7 Discussion générale

Nous aimerions premièrement souligner l'intérêt de l'approche générale présentée dans ce travail. Si la simulation de données génétiques avait déjà été effectuée dans des cadres environnementaux relativement simples (Rendine *et al.* 1986 ; Barbujani *et al.* 1995), notre équipe est sans doute parmi les premières à avoir développé cet aspect dans un cadre généraliste. L'intérêt de notre méthodologie réside dans la traduction d'informations démographiques en informations moléculaires, en tenant compte de l'influence de l'environnement. Ses avantages principaux, en comparaison de simulations classiques, sont d'une part la génération de données moléculaires – puisque les travaux antérieures ne simulaient que des fréquences alléliques – et d'autre part, l'utilisation de la théorie de la coalescence, qui permet un gain gigantesque en temps de calcul et en espace mémoire. Il sera sans aucun doute possible, dans le futur, d'effectuer des simulations individuelles selon un scénario démographique donné, mais cela n'est cependant guère réaliste actuellement dans le cas d'hypothèses complexes ou d'une résolution géographique fine. L'utilisation d'une méthode économique comme la nôtre, est le seul moyen d'étudier avec des moyens informatiques limités, des situations compliquées comme celles abordées dans ce travail.

Dans le cadre du projet "Friction", nous avons donc été amenés à défricher un vaste champ de connaissances encore très peu exploré. Nous avons ainsi dû développer de nombreux outils informatiques, comme le logiciel SPLATCHE, qui n'existaient préalablement pas. Des recherches et des expériences ont été menées dans de nombreuses directions. La thèse de Nicolas Ray (2003) présente un grand nombre de ces aspects, concernant principalement la recherche et la compilation – en version numérique – de données environnementales passées et présentes, ainsi que le développement de modèles démographiques et leur comportement dans un cadre virtuel. Pour notre part, nous avons été principalement confronté à la liaison des modules génétiques et démographiques, ainsi qu'à la pertinence des modèles et des paramètres utilisés. Nous avons également dû résoudre un certain nombre de problèmes provoqués par l'extraction et la manipulation d'un nombre gigantesque de données informatiques. Par exemple, 1 seul des 8 scénarios simulant le remplacement des Néandertaliens (chapitre 5) génère environ 100'000 fichiers de données génétiques.

Un des principaux apprentissages que nous tirons de notre participation au projet "Friction" est la nécessité d'une incrémentation graduelle de la complexité d'un modèle. Il est en effet indispensable de procéder étape par étape, afin de comprendre l'influence de chaque paramètre sur les processus étudiés. Le but final de ce travail était l'étude de l'effet des principales expansions spatiales des populations européennes sur leur structure génétique. Avant de procéder à la simulation de scénarios aussi complexes que l'arrivée des premiers hommes modernes ou la diffusion des populations néolithiques, il a été nécessaire d'étudier des situations beaucoup plus simples, afin de bien comprendre les processus impliqués. C'est pour cette raison que nous avons commencé nos recherches par la simulation d'une seule population dans un monde carré et

homogène (chapitre 3). Même si une telle aire est complètement irréaliste, elle permet de comprendre le comportement des processus génétiques dans l'espace, ces résultats étant ensuite interprétables dans un cadre réel. Nous avons ensuite augmenté à chaque fois la complexité des simulations et le nombre de paramètres considérés – par exemple en incorporant une seconde population (chapitre 4) puis une structure géographique (chapitres 5 et 6) – afin de cerner l'influence de chacun d'eux. Nous n'avons pas encore utilisé toutes les potentialités du logiciel SPLATCHE – notamment l'hétérogénéité de l'environnement et sa fluctuation au cours du temps – et notre travail peut être considéré comme un point de départ à une approche globale, sur lequel pourront se baser des études ultérieures.

Dans le cas des études exposées dans les chapitres 5 et 6, il a été nécessaire de procéder non seulement à de nombreuses recherches bibliographiques, mais également à de nombreux essais, avant d'obtenir des combinaisons de paramètres qui permettent la simulation réaliste de deux des événements démographiques majeurs qui ont affecté notre espèce sur le continent européen. Ces valeurs sont en effet tirées d'estimations archéologiques ou ethnologiques provenant de la littérature et qui ne sont souvent pas très précises (section 4.5.2). Il a donc fallu procéder à la simulation d'un très grand nombre de cas différents, afin de compenser l'incertitude qui existe autour du choix de ces paramètres. Seuls les résultats les plus représentatifs sont présentés dans ce travail. L'approche par simulation que nous utilisons ici nécessite donc une puissance informatique importante pour étudier de manière satisfaisante l'espace des résultats possibles. Nous avons heureusement pu bénéficier d'un "cluster" de 40 ordinateurs pour mener à bien nos recherches et il n'aurait pas été possible de le faire dans un temps raisonnable sans ce matériel.

La simulation de la colonisation d'une aire déjà peuplée par une nouvelle population a montré qu'il suffit de très peu d'échanges génétiques entre les nouveaux arrivants et les autochtones pour que le génome de la population invasive incorpore une grande proportion du génome de la population envahie. C'est le cas lorsque la population autochtone disparaît (comme dans les deux études présentées dans ce chapitre), mais cette "incorporation" du génome est sans doute encore plus importante lorsqu'aucune extinction n'a lieu, puisque la possibilité d'échanges génétiques est alors accrue. Cette constatation permet d'expliquer des observations "d'envahissement génétique" faites pour différents organismes, comme les truites (Bernatchez *et al.* 1995) ou les criquets (Shaw 2002). Elle permet également d'expliquer pourquoi les invasions récentes des peuplades qui ont imposé leur langage aux autochtones n'ont quasiment pas d'effet sur le patrimoine mitochondrial, comme c'est le cas en Hongrie (Semino *et al.* 2000b), dans le Caucase et en Turquie (Calafell *et al.* 1996). Lorsque une population est subdivisée, il faut qu'elle soit quasiment entièrement remplacée par une autre pour que toute trace de son patrimoine génétique disparaisse. Même un grand nombre de générations est insuffisant pour faire disparaître complètement les lignages de la population envahie par dérive génétique.

Nos simulations suggèrent d'une part, que les Néandertaliens n'ont quasiment pas contribué à notre patrimoine génétique et qu'ils peuvent donc être considérés comme faisant partie d'une branche évolutive éteinte, apparentée mais distincte de celle de notre espèce. D'autre part, il est très vraisemblable que le patrimoine génétique des chasseurs-collecteurs qui peuplaient le continent européen pendant le Paléolithique et le Mésolithique ait subsisté dans une proportion importante jusqu'à nos jours. En effet, seul un remplacement presque complet de ces chasseurs-collecteurs sur l'ensemble du continent par les descendants des premiers agriculteurs du Proche-Orient aurait pu permettre une incorporation importante de lignages proche-orientaux en Europe. Si un tel remplacement de population est possible dans certaines régions, comme l'Egée, les Balkans, l'Adriatique et l'Anatolie, les processus d'acculturation au sens large auraient largement dominé dans le reste de l'Europe (Mazurié de Keroualin 2001 ; Gallay 2004). A notre sens, s'il parait probable que les gènes des premiers agriculteurs proche-orientaux se soient répandus dans le sud-est de l'Europe lors du Néolithique, il nous semble cependant beaucoup plus improbable qu'ils soient présents majoritairement dans le reste du continent. Nous pensons que l'apport génétique global des agriculteurs proche-orientaux dans le patrimoine génétique européen est certainement minoritaire, même s'il est encore difficile de l'estimer à ce stade des connaissances.

Nous n'avons pu comparer les données génétiques virtuelles simulées qu'aux marqueurs du génome mitochondrial et à ceux de la portion non recombinante du chromosome Y. Il a été impossible de tirer des conclusions sur la base de ces derniers à cause du biais de recrutement qui les affecte. Avant d'étendre les différentes conclusions obtenues par les études présentées dans ce chapitre, il est donc nécessaire d'appliquer des analyses identiques sur d'autres parties du génome. L'étude d'un seul locus est en effet insuffisante pour tirer des conclusions définitives pour l'ensemble des populations européennes. Par exemple, la lignée masculine pourrait avoir subi une évolution différente de la lignée féminine. La multiplication des analyses sur des locus indépendants est une condition indispensable à l'établissement de conclusions irréfutables à l'aide des données génétiques. Il serait dorénavant judicieux de mettre l'accent sur le typage de marqueurs situés sur d'autres régions du génome, plutôt que de continuer à augmenter la base de données mitochondriales du continent européen dont la couverture est déjà importante. Le génome mitochondrial est énormément étudié car son haploïdie et sa grande concentration dans les cellules permettent son séquençage de manière relativement aisée, mais l'évolution extrêmement rapide des techniques de laboratoire devrait permettre dans un futur proche de disposer de bases de données importantes pour cette partie du génome. Le développement de programmes de simulation permettant la simulation de données recombinantes (Laval et Excoffier 2004), ouvrent de nouvelles perspectives à l'étude du génome nucléaire.

Nous pensons avoir suffisamment exploré l'espace des résultats possibles lors des recherches présentées dans ce travail pour bien comprendre les processus en jeu et pour présenter des conclusions robustes. Evidemment, ces dernières dépendent fortement des hypothèses sous-jacentes au modèle, notamment la compétition entre les populations et la différence de capacité de

soutien entre elles. Dans l'état actuel des connaissances, il n'est pas possible d'exclure de telles hypothèses, et le modèle que nous proposons est donc plausible pour expliquer à la fois la disparition des Néandertaliens et celle des chasseurs-collecteurs européens. Ce modèle offre d'ailleurs un surplus de réalisme par rapport à des études publiées préalablement sur le sujet, qui ne considéraient ni la subdivision des populations, ni leur dynamisme spatial (voir p. ex. Nordborg 1998). Les recherches futures permettront peut-être d'affiner ou de réfuter ce modèle, qui peut paraître relativement irréaliste sous certains aspects, notamment en ce qui concerne l'homogénéité des processus et des facteurs environnementaux.

Nous proposons donc une base de travail sur laquelle développer des travaux futurs qui pourront incorporer un surplus de réalisme. Il est cependant nécessaire qu'une augmentation de la complexité du modèle s'accompagne d'une bonne connaissance des paramètres ajoutés, afin d'éviter une croissance de l'incertitude autour des résultats et de permettre une bonne compréhension des processus en jeu. Il sera également possible d'intégrer les modules de simulations dans des procédures d'estimation de paramètres (Beaumont *et al.* 2002), afin d'estimer leurs valeurs les plus probables. Parmi les paramètres supplémentaires qui pourraient être considérés dans une étude ultérieure figure évidemment l'hétérogénéité de l'environnement. Dans le cadre de ce travail, il nous a paru suffisant de ne procéder à des simulations que dans un monde homogène. Il serait cependant très intéressant de savoir si l'utilisation d'un environnement hétérogène pour la végétation et la topographie, permettrait d'obtenir des informations complémentaires, ou au contraire ne ferait qu'obscurcir la compréhension des phénomènes. Les cartes de paléovégétation de l'Europe qui ont été récemment numérisées pour 4 périodes différentes (LGM, 20'000 BP ; Dryas récent, 11'000 BP ; début de l'Holocène, 8'000 BP et présent potentiel, ~3'000 BP: Ray et Adams 2002 ; Ray 2003: pp. 75-76) pourraient être utilisées dans ce but. De même, les cours d'eau et les côtes pourraient être pris en considération, comme voies de migration préférentielles des chasseurs-collecteurs (Anderson et Gillam 2000 ; Tolan-Smith 2003) et comme zones favorables à l'établissement des agriculteurs (Fiedel et Anthony 2003).

Nous n'avons pas considéré dans ce travail la glaciation maximum du dernier cycle glaciaire (commencé il y a environ 120'000 ans) dont le pic (LGM[1]) se situe aux alentours de 20'000 à 21'000 BP (Sherratt 1997). Pendant cette période, les températures ont drastiquement baissé en Europe et la calotte glaciaire scandinave est descendue très au sud, couvrant une partie des îles britanniques et tout le nord de l'Europe. Les grandes chaînes montagneuses étaient alors recouvertes d'épaisses couches de glace, tandis que les plaines du nord de l'Europe étaient constituées de toundra et de steppes froides. Ces conditions glaciaires ont repoussé une grande partie de la faune et de la flore plus au sud ou à l'est, dans des zones climatiques clémentes (péninsule ibérique, Italie, Balkans, Grèce, Turquie, Caucase et mer Caspienne : Huntley 1988 ; Hewitt 1996 ; Taberlet *et al.* 1998 ; Hewitt 2000 ; Willis et Whittaker 2000; Hewitt 2001). Depuis le dernier maximum glaciaire, le climat

[1] Last Glacial Maximum

s'est réchauffé progressivement – bien que de façon irrégulière – impliquant une réexpansion rapide des espèces végétales et animales vers le nord (Hewitt 1996), ainsi qu'une augmentation démographique importante des populations humaines en Europe (Bocquet-Appel et Demars 2000a). La contraction de l'aire de répartition des espèces végétales et surtout animales dans le sud et l'est de l'Europe pendant le LGM, ainsi que leur expansion post-glaciaire ont vraisemblablement été accompagnées par des modifications de l'aire de répartition des Hommes modernes, qui dépendaient fortement de ces ressources (Housley *et al.* 1997). Il a donc été proposé qu'un déplacement des chasseurs-collecteurs paléolithiques ait eu lieu en direction des zones refuges permanentes, dont les deux principales seraient : 1°) le sud-ouest de la France et la Cantabrie (Housley *et al.* 1997 ; Bocquet-Appel et Demars 2000a) ; 2°) les plaines centrales de la Russie et de l'Ukraine (Housley *et al.* 1997 ; Gribchenko et Kurenkova 1999) et éventuellement la région du Caucase (Lordkipanidze 1999). Ces contractions et expansions démographiques se sont potentiellement traduites par des modifications de la structure génétique des populations européennes (Lahr et Foley 1998 ; Roebroeks 2003). Certaines études effectuées sur les polymorphismes moléculaires des populations européennes ont d'ailleurs mentionné des traces pouvant s'apparenter à la signature d'une réexpansion post-glaciaire des populations humaines (Torroni *et al.* 1998 ; Semino *et al.* 2000a; Torroni *et al.* 2001; Underhill *et al.* 2001). Cependant, pour Barbujani et Chikhi (2000), si l'influence des colonisations post-glaciaires avait vraiment été importante, il ne serait pas possible de distinguer de gradients de fréquences continentaux entre le sud-est et le nord-ouest de l'Europe.

L'influence des contractions et des expansions sur la structure génétique des populations est un sujet qui pourrait être idéalement étudié à l'aide du programme SPLATCHE, pour autant qu'une modélisation satisfaisante du mouvement des individus vers les zones refuges soit développée. La simulation de l'abandon d'une aire par des individus engendre en effet une problématique très différente de celle de la colonisation d'une aire vide. Il serait également très intéressant de connaître l'influence que pourrait avoir ces phénomènes de contractions et de ré-expansion glaciaires sur les résultats présentés dans ce travail.

8 Conclusion générale et perspectives

Nous aimerions dans un premier temps souligner les avantages apportés par la méthodologie développée dans cette thèse, et de manière plus générale dans le cadre du projet "Friction". Nous avons montré à plusieurs reprises que l'approche que nous proposons permet de générer une certaine diversité moléculaire selon différentes hypothèses alternatives de peuplements préhistoriques. Ces signatures sont obtenues lors de situations très complexes qui peuvent difficilement être traitées analytiquement. C'est le cas, par exemple, de processus prenant place dans des populations subdivisées ou incorporant des contraintes géographiques et environnementales. Ces signatures génétiques théoriques peuvent potentiellement orienter les recherches futures, en indiquant quels sont les marqueurs – ainsi que le nombre requis – qui sont les plus aptes pour répondre à une interrogation précise. Notre méthodologie permet ainsi l'élaboration de stratégies de recherche.

L'approche présentée dans ce travail est d'autant plus intéressante que la puissance informatique ne cesse de croître et que les limites à la complexité des processus simulés sont sans cesse repoussées. De plus, cette puissance permet également de tenir compte de la stochasticité des processus démographiques et génétiques en multipliant les simulations. La méthodologie développée dans ce travail a donc porté ses fruits, en apportant un cadre théorique à l'interprétation de la structure génétique humaine, en fonction de différentes hypothèses de peuplement.

La pertinence des modèles utilisés dans les simulations est certainement le point le plus délicat de notre approche. Ils doivent en effet être suffisamment réalistes pour prendre en compte les principaux éléments décrivant les situations désirées, mais tout de même assez simples pour que les processus en jeu puissent être compris. Par définition, un modèle ne sera jamais conforme à la réalité, puisqu'il ne sera jamais possible de simuler exactement l'histoire des populations humaines (ou d'autres organismes) telle qu'elle s'est déroulée. En revanche, si des hypothèses suffisamment différenciées sont proposées, et qu'il est possible de les modéliser de façon raisonnable, alors leurs signatures génétiques respectives sont potentiellement identifiables au moyen de simulations. Nous avons notamment montré que les modèles de diffusion démique ou de diffusion culturelle du Néolithique peuvent être différenciés grâce à leurs diversités moléculaires (chapitre 6). De même, les structures génétiques attendues sous les hypothèses d'une origine unique ou multiple de l'Homme moderne peuvent également être distinguées (Ray *et al.* 2004). Dans les deux cas, il s'agit cependant d'hypothèses extrêmes et opposées, et il est beaucoup plus difficile d'évaluer la pertinence des situations intermédiaires à l'aide de notre méthodologie. Il s'agit d'ailleurs de la principale faiblesse de notre approche, en l'état. Elle est en effet extrêmement utile pour différencier qualitativement les données moléculaires en fonction des différents scénarios et ainsi comprendre l'effet des processus en jeu. En revanche, elle manque de puissance pour permettre la comparaison quantitative des scénarios en fonction des données moléculaires réelles, excepté dans certains cas (comme dans la recherche présentée dans le chapitre 5). Gageons que l'utilisation d'une méthode bayesienne d'estimation de paramètres (voir p. ex. : Beaumont *et al.* 2002), combinée à notre

approche, devrait permettre dans le futur une évaluation plus précise des scénarios, même s'il existera toujours une limite dans la reconstitution des données réelles.

Les différentes recherches présentées dans ce travail ont néanmoins permis de tirer plusieurs conclusions au sujet des populations humaines. Nous ne reviendrons que brièvement sur ces différents points puisqu'ils ont déjà été largement abordés dans cette thèse.

- Une population subdivisée ayant passé par une expansion spatiale et démographique présente une diversité moléculaire différente en fonction du nombre de migrants échangés entre les sous-populations qui la composent (le paramètre *Nm*). Cette observation permet d'expliquer, par une simple différence de densité, les différences observées dans la diversité moléculaire des populations de chasseurs-collecteurs contemporains et dans les populations post-néolithiques (section 3.2).
- La trace d'une expansion paléolithique de la lignée mâle en Europe peut être indécelable dans les distributions "mismatch" établies avec des données de type SNP (section 3.3).
- Sous l'hypothèse d'une vague de migration des hommes modernes depuis le sud-ouest de l'Asie, le patrimoine mitochondrial européen actuel ne peut résulter que d'une hybridation extrêmement faible, voire nulle, avec les Néandertaliens. Par conséquent, ces derniers ne font pas partie de nos ancêtres directs, mais appartiennent à une espèce distincte de la nôtre, dont la lignée s'est éteinte (chapitre 5).
- Des gradients de fréquences alléliques entre le Proche-Orient et le nord-ouest de l'Europe peuvent avoir été générés aussi bien par la vague de migration des premiers Hommes modernes, il y a 40'000 ans, que par celle des premiers agriculteurs, il y a 10'000 ans. L'observation de tels gradients n'est donc pas une preuve de la diffusion démique du Néolithique (chapitre 6).
- Le génome mitochondrial européen est compatible avec une forte contribution des chasseurs-collecteurs paléolithiques dans le patrimoine génétique féminin, lors de la transition néolithique (section 4.5.4 et chapitre 6). Cette observation va à l'encontre de certaines estimations faites pour le chromosome Y, qui semble avoir été affecté de façon plus importante par les agriculteurs proche-orientaux. Si cette hypothèse venait à être confirmée par l'étude d'autres locus, elle pourrait être expliquée par une transmission des techniques agropastorales davantage par voie masculine que par voie féminine.

Le choix des types de polymorphismes étudiés (séquence, SNP, RFLP, STR, allozyme) est d'une très grande importance lors de la comparaison entre différents systèmes génétiques. Par exemple, la différence dans les distributions "mismatch" observées en Europe pour les hommes et les femmes peut être due au type de données étudiées et non à une différence démographique entre les deux lignées. En effet, le système génétique spécifique à la lignée mâle – la partie non-recombinante du chromosome Y (MSY) – a été typé principalement avec des SNPs et des STRs, alors que celui spécifique à la lignée féminine – le génome mitochondrial – est principalement étudié par des séquences d'ADN. Or, les échantillons composés de SNPs et de STRs sont sujets à un biais de recrutement beaucoup plus important que ceux qui composés de séquences. Ce biais

implique que les mutations "récentes" sont sous-représentées parmi les sites étudiés, révélant une image différente des données. Par conséquent, la comparaison des systèmes liés au sexe peut donner lieu à des interprétations erronées si ce facteur n'est pas pris en compte. Ce biais de recrutement explique également pourquoi des gradients de fréquences alléliques sont observés beaucoup moins souvent avec le génome mitochondrial qu'avec d'autres sytèmes. Nous avons en effet montré que l'observation d'un gradient à la suite d'une vague de migration est tributaire de l'âge de la mutation étudiée, la présence de gradients étant d'autant plus élevée que le biais de recrutement en faveur des allèles fréquents est fort (chapitre 6).

Comme nous l'avons déjà mentionné, ce travail constitue une première étape dans une approche plus globale visant à fournir un cadre théorique à l'interprétation de données réelles. Il en émerge donc de nombreuses perspectives. Nous ne reviendrons pas sur celles dégagées par nos études sur le peuplement de l'Europe, puisque nous les avons déjà décrites dans le chapitre précédent (chapitre 7). En revanche, il serait extrêmement intéressant d'étudier l'influence que pourraient avoir de nouveaux paramètres sur la signature génétique des populations humaines ou d'autres organismes terrestres :

- Les migrations à longue distance jouent un rôle prépondérant à la fois dans la vitesse d'une expansion (Nichols et Hewitt 1994) et dans la diffusion des génomes. En effet, les lignages transportés par ces migrations vont contribuer plus fortement au patrimoine génétique final, puisque leur arrivée dans des zones vierges va leur permettre une diffusion très rapide (Hewitt 1996). Ces mouvements à longue distance ont certainement été importants chez l'Homme (Sokal 1991b ; Langaney *et al.* 1992).

- La prise en compte de l'hétérogénéité de la végétation, ainsi que de la topologie du terrain, permettrait sans doute de révéler des routes de migration préférentielles et leur influence sur la structure génétique finale. Cependant, l'utilisation d'un environnement hétérogène – par exemple en utilisant des cartes de paléovégétation (Adams et Faure 1997 ; Ray et Adams 2001) – reste délicat, car il est nécessaire d'attribuer des valeurs de capacité de soutien à tous les types de végétation (Ray 2003 : pp. 77-80 ; Ray *et al.* 2004).

- Les fluctuations du climat au cours du temps impliquent des transformations dans les types de végétation, ainsi que des variations du niveau des mers. Non seulement ces changements modifient l'aire de répartition des espèces et l'accessibilité à certaines régions – comme par exemple l'accès aux Amériques par le Détroit de Bering (Fiedel 1992 ; Crawford 1998) – mais elles peuvent également catalyser l'apparition d'innovations (culturelles ou techniques) en créant les conditions favorables à leur éclosion. Par exemple, pendant les périodes froides, le Moyen-Orient devient une sorte de cul-de-sac où les populations (humaines et animales) du nord de l'Afrique et celles de l'Europe peuvent se rencontrer[1]. Ces conditions climatiques ont peut-être favorisé la réunion des

---

[1] Pendant les périodes climatiques froides, la limite de la calotte polaire européenne est très basse, obligeant les populations à se déplacer vers le sud. Le Sahara étant à son extension maximum pendant ces mêmes périodes froides, les populations nord-africaines vont, quant à elles, se déplacer vers le Moyen-Orient (Sherratt 1997 ; Lahr et Foley 1998).

différents facteurs économiques et culturels, qui ont permis à plusieurs reprises l'émergence de nouvelles technologies dans cette région, notamment les premières composantes du Néolithique européen (Sherratt 1997).

- La simulation de la recombinaison entre locus diploïdes permettrait d'étudier l'effet de différents événements démographiques sur le déséquilibre de liaison[1]. Cet aspect offre, en effet, des perspectives prometteuses. Kaessmann (2002) a, par exemple, montré que les dernières populations de chasseurs-collecteurs européens (Saamis et Evenkis) présentent un déséquilibre de liaison plus fort que celui des autres populations européennes.

Notre approche peut évidemment être utilisée dans le cadre d'une large variété de questions, touchant soit les populations humaines, soit d'autres organismes. Elle pourrait, par exemple, apporter des éléments nouveaux à la compréhension du peuplement de la Polynésie, en permettant la simulation des différentes hypothèses proposées (p. ex. : Oppenheimer et Richards 2001). Par ailleurs, une version de SPLATCHE permettant de modéliser le déplacement d'organismes dans des cours d'eau est également en cours de développement dans le cadre de la thèse de Samuel Neuenschwander (in prep.), à l'Université de Berne. Cette version modifiée de SPLATCHE permettra notamment d'aborder les questions relatives à la recolonisation des plans d'eaux après les phases de glaciation.

Comme nous l'avons déjà souligné à plusieurs reprises, la méthodologie présentée dans ce travail pourrait certainement tirer d'énormes bénéfices de sa combinaison avec des méthodes d'estimations bayesiennes, qui donnerait une évalutation plus précise des scénarios simulés et une estimation des paramètres démographiques des populations réelles. Ce type de méthodologie requiert cependant des capacités informatiques très importantes puisqu'elle nécessite des millions de simulations (Beaumont et Rannala 2004 ; Hamilton *et al.* 2004).

---

[1] Le "déséquilibre de liaison" désigne l'association non-aléatoire d'allèles pris à des locus séparés, dans une population (Lewontin et Kojima 1960). Ce phénomène peut être la conséquence de la sélection, de la dérive génétique, de la parenté entre individus ou du flux génétique entre populations (Lewontin 1988).

9 Annexes

ANNEXE 1 Manuel d'utilisation de SPLATCHE

Cette annexe est une reproduction du manuel d'utilisateur qui accompagne la version "publique" du logiciel "SPLATCHE", disponible à l'adresse www.cmpg.unibe.ch/software/SPLATCHE.

{Page suivante}

SPLATCHE: USER MANUAL

1 Introduction

The goal of this user manual is to describe the technical aspects of the software SPLATCHE (version 1.0). This manual complements the article from Currat, Ray and Excoffier, published in *Molecular Ecology Notes* (Currat *et al.* 2004). Further details on the methodology can also be found in Ray (2003a) and Currat (in prep.).

2 Contents

3 Demographic and spatial expansion module

3.1 Principles

The demographic and spatial expansion module allows to simulate a demographic and spatial expansion from one or many initial populations. The simulation uses discrete time and space. The unit of time is the generation, while the unit of the 2D space is a cell, also called a deme. Each deme has the same size and can be considered as a homogeneous subpopulation. The spatial model used in SPLATCHE is the 2D stepping-stone model (Kimura & Weiss 1964), which defines a regularly spaced array of demes. Each deme undergoes an independent population growth and can exchange emigrants with its four direct neighboring demes.
Each deme is also considered as a sub-unit of the environment. The environment can influence the local demography through its carrying capacity (maximum number of individuals) and its friction (facility to migrate through). These two environmental characteristics can be defined for the entire array of demes through the input of maps. Variations, through time, of carrying capacity and/or friction values are also possible.

3.2 Available demographic models

The logistic population growth of each deme follows a standard logistic curve, of the form

$$N_{t+1} = N_t\left(1 + r\frac{K - N_t}{K}\right),$$

where K is the carrying capacity, and r is the growth rate.
For the migration part of the demography, three models are available in SPLATCH:

Model 1. Migration model with even number of emigrants
The number of emigrants M from a deme is computed, for each generation, as
$M = mN_t$, where m is the migration rate, and N_t is the population density of the deme at generation t. The number of emigrants M_i in any of the four directions is then computed as

$$M_i = floor\left(mN_t \cdot \frac{1}{F_i \cdot \sum_{j=1}^{4}\frac{1}{F_j}}\right),$$

where F_i is the friction of the deme in direction i (north, south, east or west), and $floor$ means that the fractional part of the number is truncated. This model always gives a total number of emigrants which is a multiple of four.

Model 2. Migration model with absolute number of emigrants
Same as Model 1, but the fractional part of M_i is not truncated. Instead, a multinomial distribution is used to split M emigrants to the neighboring demes (see Ray 2003a). This ensures that there are always M emigrants that are sent. The drawback of this technique is that it requires the drawing of random numbers, which increases the time required for a simulation.

Model 3. Stochastic migration model with absolute number of emigrants
Same as Model 2, but the number of emigrants M varies stochastically as a Poisson variable centered around $N_t m$.

3.3 General Settings panel

The *General Settings* panel is the primary panel to set the demographic parameters and to launch a demographic simulation. A screenshot of this panel is shown in Figure 3.1. A description of each component of this panel is given in the following sub-chapters.

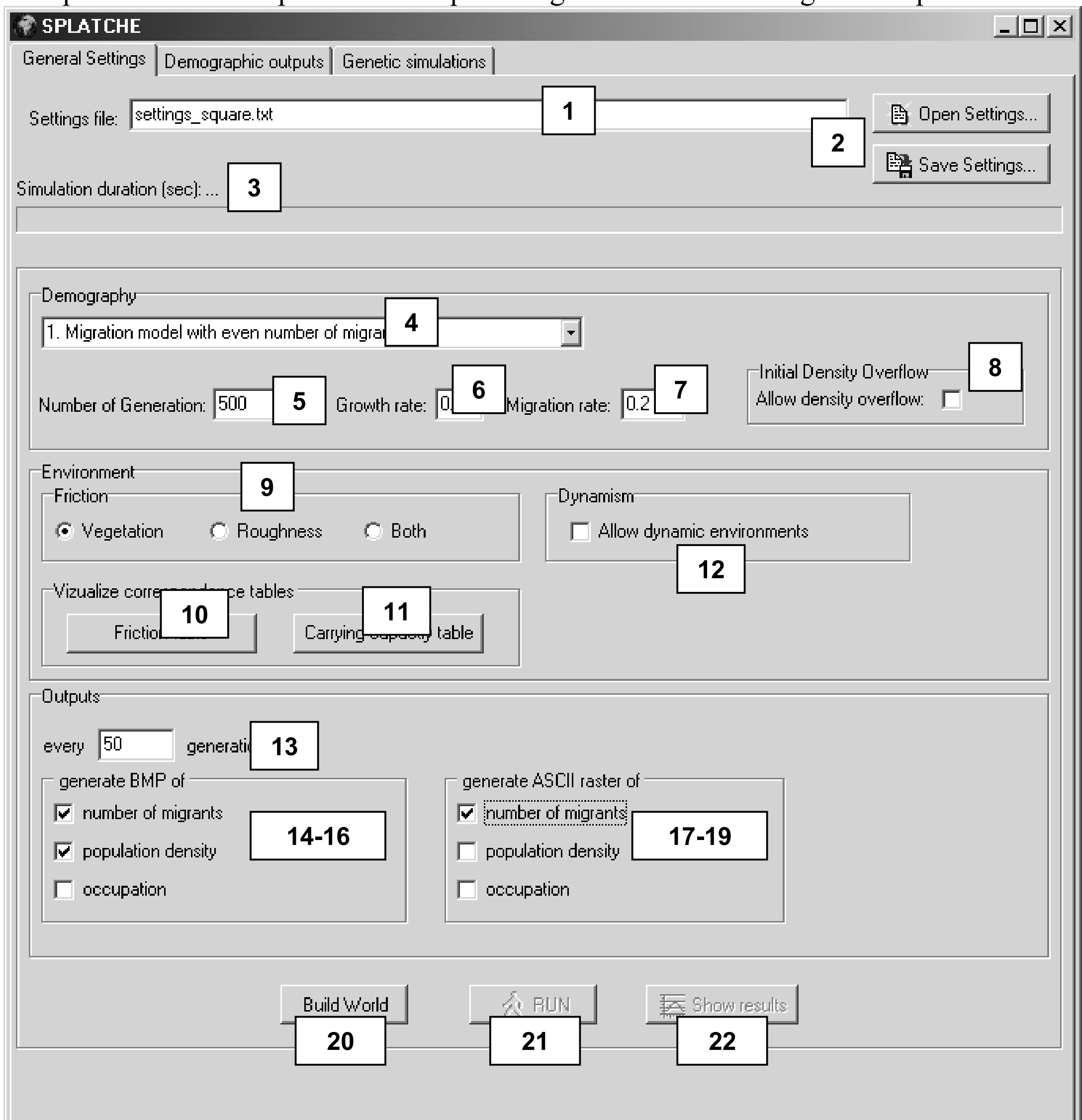


Figure 3.1. General Setting panel. The numbers correspond to a description in the text.

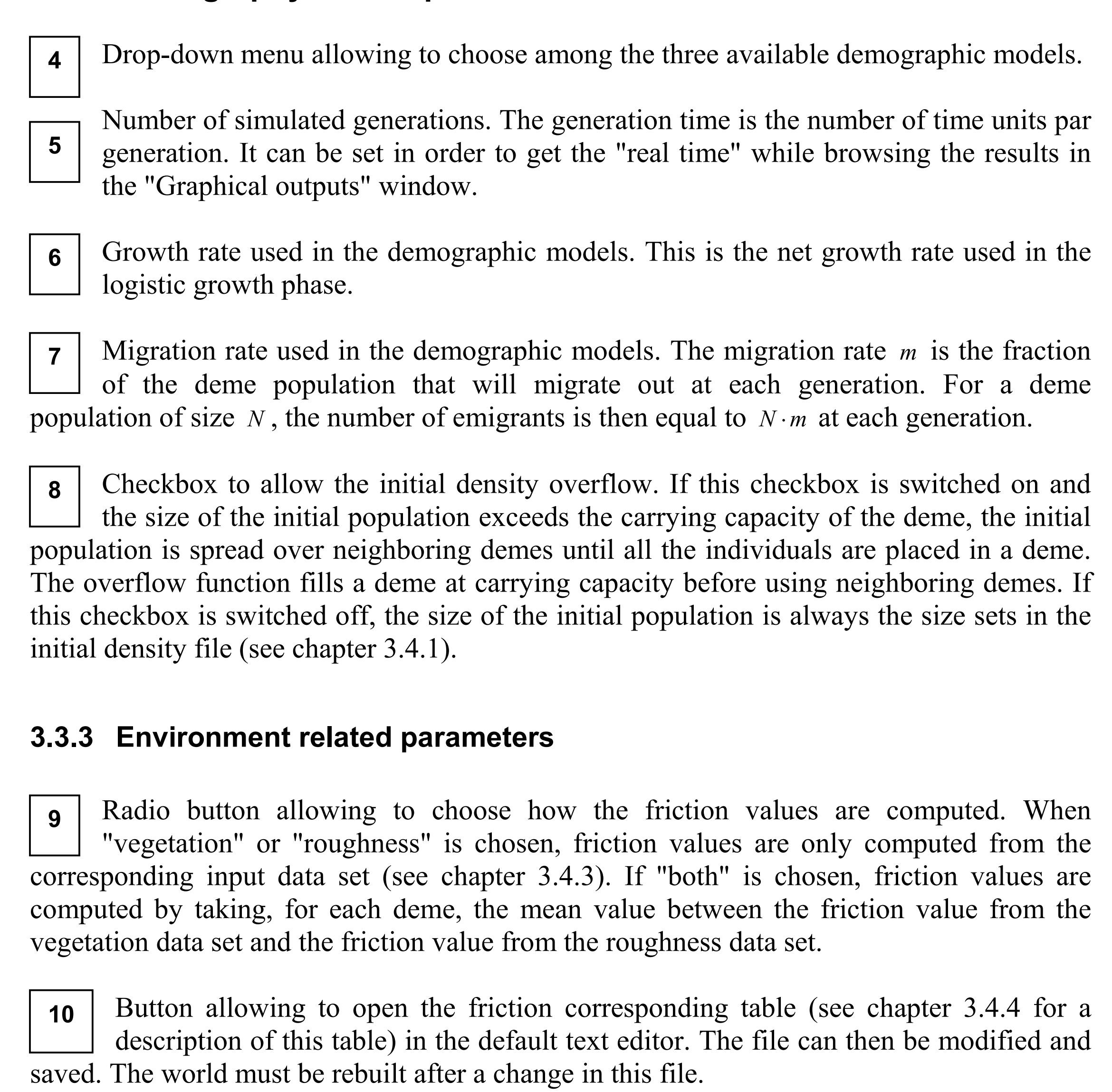
3.3.1 General

1 Settings file name: location of the settings file (*.txt). See chapter 3.4.2 for the full description of a settings file.

2 Buttons allowing to open a settings file or to save a settings file.

3 Progress bar showing the remaining computation time of a current simulation. The duration of a simulation (in seconds) is also given at the end of the computation.

3.3.2 Demography related parameters



4 Drop-down menu allowing to choose among the three available demographic models.

5 Number of simulated generations. The generation time is the number of time units par generation. It can be set in order to get the "real time" while browsing the results in the "Graphical outputs" window.

6 Growth rate used in the demographic models. This is the net growth rate used in the logistic growth phase.

7 Migration rate used in the demographic models. The migration rate m is the fraction of the deme population that will migrate out at each generation. For a deme population of size N, the number of emigrants is then equal to $N \cdot m$ at each generation.

8 Checkbox to allow the initial density overflow. If this checkbox is switched on and the size of the initial population exceeds the carrying capacity of the deme, the initial population is spread over neighboring demes until all the individuals are placed in a deme. The overflow function fills a deme at carrying capacity before using neighboring demes. If this checkbox is switched off, the size of the initial population is always the size sets in the initial density file (see chapter 3.4.1).

3.3.3 Environment related parameters

9 Radio button allowing to choose how the friction values are computed. When "vegetation" or "roughness" is chosen, friction values are only computed from the corresponding input data set (see chapter 3.4.3). If "both" is chosen, friction values are computed by taking, for each deme, the mean value between the friction value from the vegetation data set and the friction value from the roughness data set.

10 Button allowing to open the friction corresponding table (see chapter 3.4.4 for a description of this table) in the default text editor. The file can then be modified and saved. The world must be rebuilt after a change in this file.

11 Button allowing to open the carrying capacity corresponding table (see chapter 3.4.4 for a description of this table) in the default text editor. The file can then be modified and saved. The world must be rebuilt after a change in this file.

12 CheckBox allowing a dynamic simulation (see chapter 3.4.4). The world must be rebuilt after a change in this checkbox.

3.3.4 Output parameters

Some output parameters are placed in this panel, because they need to be set prior to a simulation, if one wants to automatically generate these outputs during the simulation. These outputs are a temporal series of graphical representations of the state of a demographic parameter (number of emigrants, population densities, or occupation). Windows Bitmaps (BMP) or ASCII raster can be generated. The output files are placed in two folders (called respectively, "BMP" and "ASCII") which are created in the same folder

than the setting file. The filename of each output file is composed by the name of the demographic variable followed by the number of generation at which it has been created.

13 Number of generations between each output files. Beside the outputs for the intermediate states, a series has always outputs for the initial and the final state of the simulation.

14-16 Checkboxes for the generation of BMP files.

17-19 Checkboxes for the generation of ASCII raster files.

3.3.5 Main buttons

20 Button to build a world. It is during a building process that memory space is allocated, and that carrying capacity and friction values are computed for each deme

21 Button to launch a simulation. If this button is grayed out, it means that the world needs to be built or rebuilt.

22 Button to show the graphical output window.

3.4 Input files

3.4.1 Initial density and origin location

A file, called "dens_init.txt" in the examples, is used to specify the place of origin of the simulated population. This file contains a first line of legend and a second line defining the population source. This second line consists of 5 fields separated by "tab" or "space" character:
1. Name of the source population.
2. Size of the source population, in number of effective haploid individuals.
3. & **4**. Geographic coordinates of the population source (latitude and longitude). SPLATCHE will determine itself in which particular deme corresponds the coordinates of the population. Coordinates must belong to the geographical surface defined in the header of the environmental files. Coordinates do not need to be in a particular units (e.g. decimal degrees), but they needs to be in the same units that the coordinates defined in the header of the environmental files.
5. Resize parameter: it is the size of the population source before the beginning of the expansion. This parameter is used only for genetic simulations. If this parameter is set to 0, then the size of the population source before the onset of the expansion is regarded as being equal to the initial size (parameter 2.).

Example of initial density file:

```
#Name  #Size  #Lat   #Long  #Resize
middle 100    -1     13     0
```

3.4.2 Settings file

All the parameters can be defined using the graphical interface of SPLATCHE. However, it is possible to save a group of parameters into a file, and thus of being able to recover them thereafter. Only the graphical parameters are not contained in the settings files. An example of settings files aisre provided with SPLATCHE: “settings_square.txt”, with the corresponding data files in the folder called "dataSets_square". The example file is a simple square world constituted by 50x50 demes (see Ray *et al.* 2003a).

The setting file is composed of 29 parameters. An example, corresponding to "settings_square.txt", is given below. Each line starts with the value of the parameter, followed by a blank, a double slash, and then the description of the parameter.

```
./dataSets_square/dens_init.txt //pop source file
./dataSets_square/simplesquare.asc //vegetation file
./dataSets_square/simplesquare.asc //roughness file
./dataSets_square/Dynamic_K.txt //Conversion table Vegetation->K
./dataSets_square/Dynamic_F.txt //Conversion table Vegetation->F
3 //demographic model (1-3)
700 //number of generations
0.10 //growth rate
0.20 //migration rate
0 //allow Initial Density overflow? (0/1)
1 //static or dynamic environment? (0/1)
0 //choice of friction type (0:vegetation,1:roughness topography,2:both)
./dataSets_square/genes_middle.sam //original genetic sample file
1 //number of genetic simulations
10000 //maximum number of simulated generations
0 //Genetic Data Type (0:DNA,1:RFLP,2:MICROSAT,3:STANDARD)
300 //number of linked loci
0.001 //total mutation rate
0.33 //fraction of substitutions being transitions for DNA
0 //Gamma A for DNA mutation variation
0 //number of Categories for DNA mutation variation
0 //Range Constrainst for microsatellite
0 //generate Arlequin file, Paup files or both (0/1/2)
0 //generate migration BMP
0 //generate density BMP
1 //generate occupation BMP
0 //generate migration ASCII
1 //generate density ASCII
0 //generate occupation ASCII
```

3.4.3 ASCII format for environmental data

The environmental datasets that can be loaded into SPLATCH must be in ASCII raster format. Two different datasets can be loaded. The first one is the "vegetation" dataset, defining to what type (category) of vegetation belongs each deme. The second dataset is the "roughness" dataset, defining continuous friction values, such as friction computed from topography.

This format of the environmental dataset is composed of a header (first six lines) containing information on the file, then a matrix of values in rows and columns.

The header information is as follow:

ncols : number of columns
nrows : number of rows
xllcorner : longitude coordinate of the lower-left deme
yllcorner : latitude coordinate of the lower-left deme
cellsize : width of a deme (cell size), in same units than the coordinates

NODATA_value : value indicating than a deme must not be considered (like sea)

Example of an environmental dataset

```
ncols         88
nrows         91
xllcorner     -19.845388
yllcorner     -36.897187
cellsize      0.83
NODATA_value  -9999
-9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 7 7 7 ...
-9999 -9999 -9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7
-9999 -9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7 7 7 7
-9999 -9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7 7 7 7
-9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7 7 7 7 7 7 7
-9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7 7 7 7 7 7 7
...
```

3.4.4 Dynamic simulations and conversion tables to obtain K and F

It is possible in SPLATCHE to do dynamic simulations. A dynamic simulation allows variation of carrying capacity and/or friction value at different time during the course of a simulation. In order to set at what time the changes occur, different files are needed

The two main files, which are set through the settings files, are typically called "Dynamic_K.txt" and "Dynamic_F.txt". On the first line of each of this file appears the number of changes during a simulation. Then each line (one per change) is composed by the time of change (in generations), the filename of the corresponding table (see below), and an arbitrary description. The three components of each line must be separated by a blank space. For a non-dynamic simulation, only the first filename is considered, regardless of the number indicated on the first line.

Example of "Dynamic_K.txt" file:

```
2
0 ./dataSets_africa/veg2K.txt vegetation at time 0
500 ./dataSets_africa/veg2K_500.txt doubling of vegetation at time 500
```

Each file name must targets to a valid "corresponding table" that makes the link between a particular vegetation category and a carrying capacity (or friction) value. A corresponding table is composed of a vegetation category number, followed by a carrying capacity (or friction) value, and by a description. The vegetation category numbers must correspond to the numbers found in the input "vegetation" dataset (see previous chapter).

Example of "veg2K.txt" file:

```
1      200    Tropical_rainforest
2      200    Monsoon_or_dry_forest
3      500    Tropical_woodland
4      500    Tropical_scrub
5      100    Tropical_semi_desert
6      1000   Tropical_grassland
7      50     Tropical_extreme_desert (50)
8      1000   Savanna
9      200    Broadleaved_temperate_evergreen_forest
10     200    Montane_tropical_forest
11     500    Open_boreal_woodlands
12     500    Semi_arid_temperate_woodland
```

By having several corresponding tables for the carrying capacity and/or the friction values, it is then possible to simulate a change in the environment through time.

3.5 Graphical outputs window

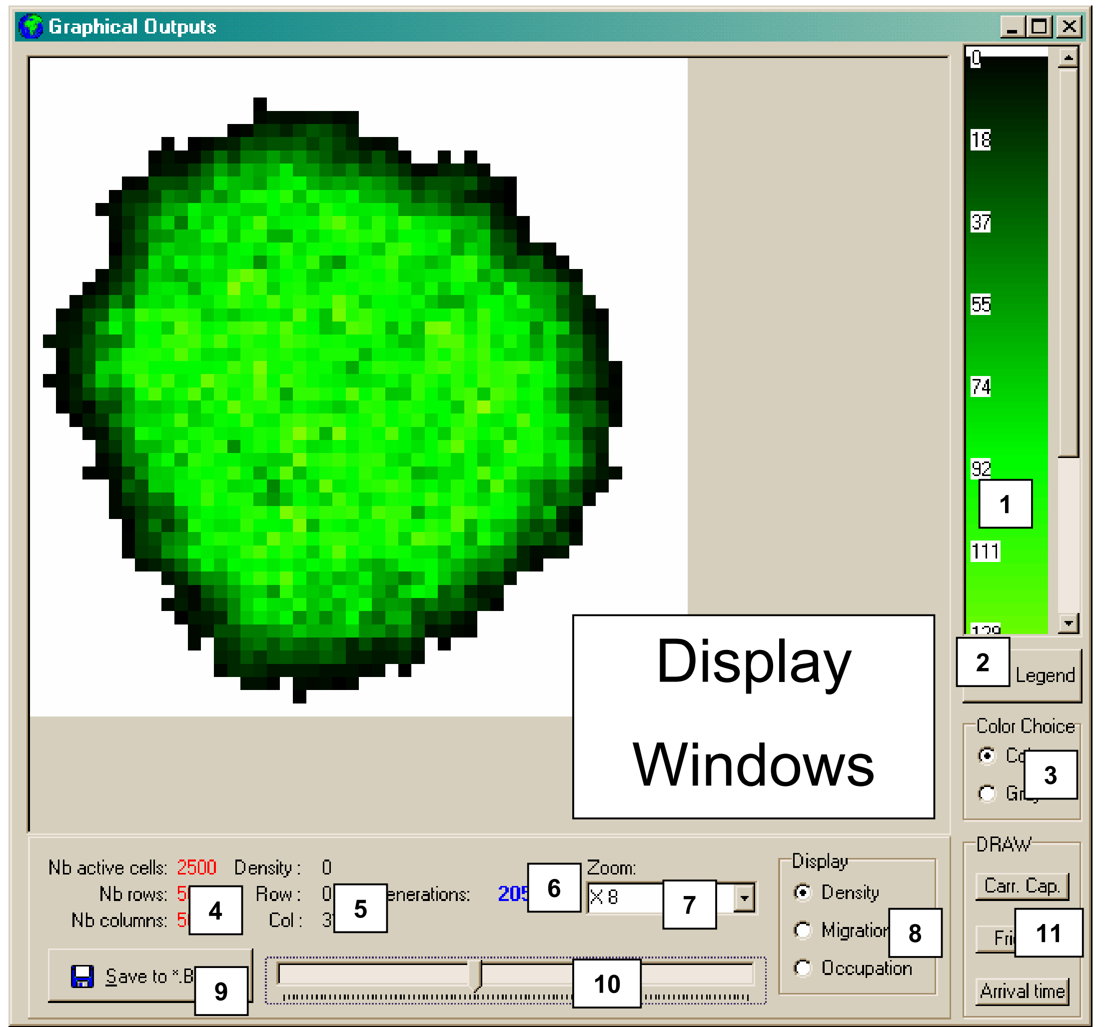


Figure 3.2. Graphical Outputs panel. The numbers correspond to a description in the text.

1 Legend for the current display.

2 Buttons allowing to save the legend as a bitmap.

3 Radio button for the choice of color or shades of gray display.

4 Information on the number of active cells (cells having information for the vegetation), the number of rows and the number of columns.

5 Information on the density, the number of rows and the number of columns when the mouse cursor is over a particular deme.

6 Number of generations for the current display.

7 Zoom for the current display.

8 Radio button to choose from displaying the density, the number of emigrants, or the occupation (black if occupied).

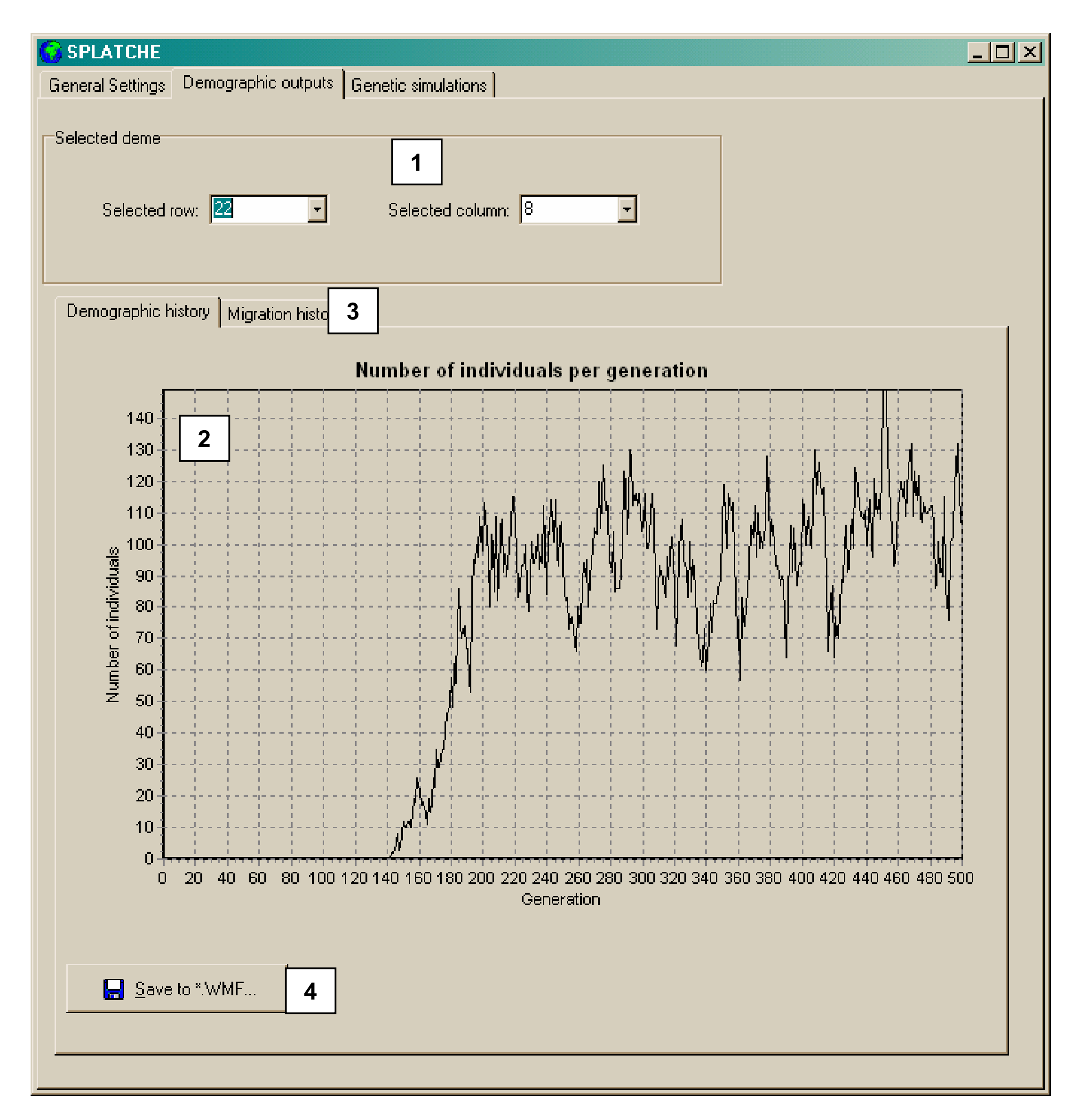
9 Button allowing to save the current display as a bitmap.

10 Cursor allowing to change the current generation, and the display at the chosen generation.

11 Buttons allowing to display the initial (at generation 0) carrying capacity map, the initial friction map, and the proportional arrival time in each deme.

3.6 Demographic outputs window

This window allows to explore the demographic database that has been generated through a simulation.



1 Selectors allowing to change the row and the column, which will select the correct deme and display the history of the number of individuals (density).

2 Graph showing the history of the number of individuals (density) for the selected deme. It is possible to zoom in and out in the graph by drawing rectangles with the mouse cursor (left button down).

3 Second panel showing the histories of the number of emigrants in the four directions.

4 Button allowing to save the graph in Windows Metafile format.

4 Genetic module

4.1 Principles

Genetic simulations are always preceded by a demographic simulation. Indeed, a genetic simulation uses the demographic information stored in the data base generated during the demographic phase. The genetic phase is based on the "coalescent theory", initially described by Kingman (1982a, 1982b) and developed in other papers (Ewens 1990; Hudson 1990a; Donelly et Tavaré 1995). This theory allows the reconstruction of the genealogy of a series of sampled genes until their most recent common ancestor (MRCA). For neutral genes, the genealogy essentially depends on the demographic factors that have influenced the history of the populations from whom the genes are drawn. The implementation of the coalescent theory is a modified version of SIMCOAL (Excoffier et al. 2000). The principal difference with SIMCOAL is that the demographic information used by genetic simulations do not come anymore from the "migration matrix" and "historical events", but from the data base generated during the demographic simulation.

The genetic simulation itself follows the procedure described in Excoffier *et al.* (2000) and consists in two phases:

1°) Reconstruction of the genealogy:
The reconstruction of the genealogy is independent of the mutational process. Basically, a number *n* of genes is chosen. These genes are only identified by their number and they have no genetic variability during this first phase. All the *n* genes are associated with a geographic position in the virtual world where the demography is simulated. These genes could belong to different demes in the world. Then, going backward in time, the genealogy of these genes is reconstructed until their most recent common ancestor (MRCA) in the following way:
Going backward in time, at each generation, two events can occur:

- ***Coalescent event:*** if at least two genes are on the same deme, they have a probability to have a common ancestor at the preceding generation (a coalescent event). This probability depends on the population size of the deme where the genes are located. Each pair of genes has a probability $1/ N_i$ of coalescence (if N_i is the number of haploid individual in the deme *i*). If there are n_i genes on the deme then the probability of one coalescent event becomes $n_i (n_i -1)/ 2N_i$. Only one coalescent event is allowed per deme and per generation (see Ray *et al.* 2003a) for a discussion about this assumption).

- ***Migration:*** Each gene could have arrived with an immigrant from a different deme. When going back in time, it means that the gene could leave the current deme with the immigrant. So, the probability of migration from a deme *i* to a deme *j* for a gene depends on the number of individuals that have arrived from deme *j* to deme *i* at this generation. For each gene belonging to the deme *i*, the probability of migration from deme *j* is equal to m_{ji}/N_i where m_{ji} is the number of immigrants from deme *j* to deme *i* during the demographic phase.

All the deme sizes and the numbers of immigrant between demes are taken from the database generated during the demographic simulation.
2°) Generation of the genetic diversity:
The second phase of a genetic simulation consists in generating the genetic diversity of the samples. This operation is done in adding mutations independently on all branches of the

genealogy assuming a uniform and constant Poisson process. At the end of this process all the sampled genes have a specific genetic identity. The genetic process is entirely stochastic, so many genetic simulations have to be performed for each demographic simulation in order to obtain meaningful statistics. We recommend at least 1'000 simulations per demographic scenario.
The coalescent backward approach does not generate the history of the whole population, but only that of sampled genes and their ancestors. Thus this approach is much less demanding in terms of memory and computing time. That allows the simulation of complex demographic scenarios within a very broad geographical and temporal framework.

4.2 Settings panel

Various parameters must be defined before launching a genetic simulation. The number of parameters can be seen in Figure 4.1.

4.2.1 General

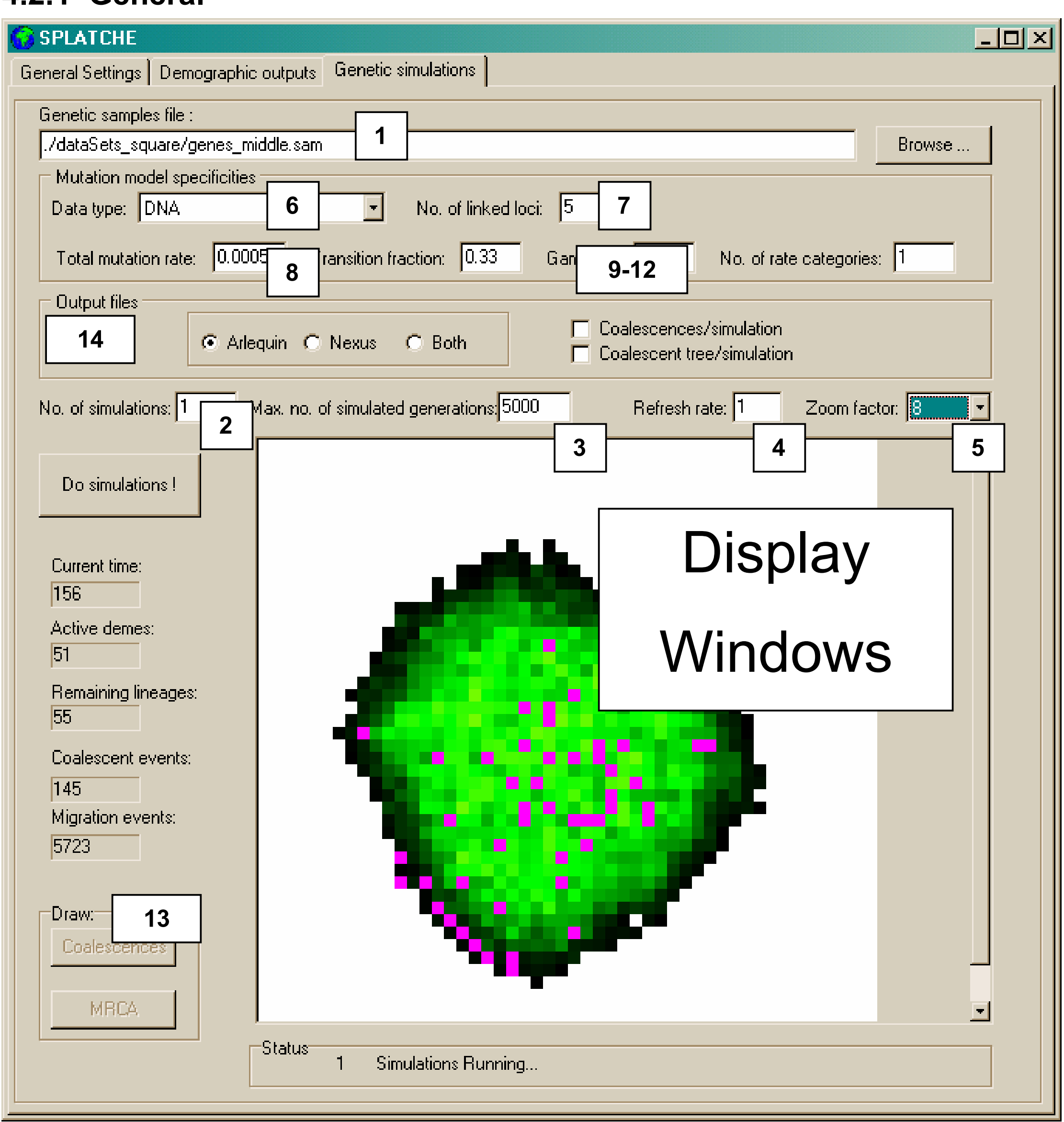


Figure 4.1 Genetic module panel. Demes where at least one gene is present appear in violet.

1 Sample file name: location of the *.sam file.

2 Number of simulations to be carried out

3 Maximum of generations after which the process stop if the genealogy has not been correctly reconstructed.

4 Refresh rate: generation numbers after which the display window is updated.

5 Zoom factor of the display window.

4.2.2 Mutation model specificities

-For all kind of data:

6 Type of genetic data to be generated. It could be DNA, RFLP, Microsatellite or Standard. See "Genetic data type" Section for more details.

7 Number of fully linked loci to simulate. It corresponds to the sequence length for DNA.

8 Mutation rate per generation for all loci taken together.

-Specific to DNA:

9 Transition bias: percentage of substitutions that are transitions.

10 *Gamma a: amount* of heterogeneity in mutation rates along the sequence according to either a discrete or continuous gamma distribution.

11 Number of categories for DNA mutation variation.

-Specific to Microsatellite:

12 Range constraint: minimum and maximum size for microsatellite.

4.2.3 Genetic data

Different types of molecular data could be generated (RFLP, DNA, Microsatellites and Standard), each with its own specificities:

-RFLP data: Only a pure 2-allele model is implemented. Several fully linked RFLP loci can be simulated, assuming a homogeneous mutational process over all loci. A finite-sites model is used, and mutations can hit the same site several times, switching the RFLP site on and off. We thus assume that there is the same probability for a site loss or for a site gain.

-Microsatellite data: We have implemented a pure stepwise mutation model (SMM) with or without constraint on the total size of the microsatellite. Several fully linked microsatellite loci can be simulated under the same mutation model constraints. The output for each loci is listed as a number of repeat, having started arbitrarily at 10,000 repeats. The number of repeats for each gene should thus be centered around that value.

-DNA sequence data: We have implemented here several simple finite-sites mutational models. The user can specify the percentage of substitutions that are transitions (the transition bias), the amount of heterogeneity in mutation rates along a DNA sequence according to either a discrete or continuous Gamma distribution. We can therefore simulate DNA sequences under a Jukes and Cantor model (Jukes et Cantor 1969) or under a Kimura-2-parameter model (Kimura 1980), with or without Gamma correction for heterogeneity of mutation rates (Jin et Nei 1990). Other mutation models that depend on the nucleotide composition of the sequence were not considered here, because of their complexity and because they require specifying many additional parameters, like the mutation transition matrix and the equilibrium nucleotide composition.

-Standard data: Following the definition given in Arlequin User Manual (Schneider *et al.* 2000a), this type defines data for which the molecular basis is not particularly defined, such as mere allele frequences. The comparison between alleles is done at each locus. For each locus, the alleles could be either similar or different.

4.3 Input files

4.3.1 Genetic samples

A file with the extansion ".sam" allows to specify the localization of the population sampled, as well as the number of genes sampled in each population.
On the first line of this file, the user can specify the number (integer) of population sampled. The second line is reserved for the legends. Then, each line defines a sample with 4 fields separated by "tab" our "space" character.
1. Name of the population from which the sample has been drawn.
2. Number of genes belonging to that sample.
3. & **4**. Geographic location of the population. (latitude and longitude). SPLATCHE will determine automatically in which particular deme falls the coordinates of the population. The coordinates must belong to the geographical surface defined in the header file.

Example of a genetic input file (.sam) for 6 samples in Africa:

```
6
#Name        #Size  #Lat   #Long
sample1      30     20     20
sample2      25     20     0
sample3      28     0      20
sample4      32     -20    20
sample5      30     -30    25
sample6      40     5      40
```

4.4 Output files

13 **14** on Figure 4.1.

Various kinds of genetic output files can be generated by SPLATCHE:

4.4.1 Arlequin files

The genetic data generated by one simulation are directly output in an ARLEQUIN project file, with the extension "*.*arp*". This file format allowing one to compute the data using the ARLEQUIN software in order to obtain different statistics, see ARLEQUIN manual (Schneider *et al.* 200a) for more details. If more than one simulation is performed using one demographic simulation (which is usually the case) then an ARLEQUIN batch file (with extension "*.*arb*") is also generated, listing all simulated files, and allowing one to compute statistics on the whole set of simulated files. Note also that the ARLEQUIN software has a file conversion utility for exporting input data files into several other format like BIOSYS, PHYLIP, or GENEPOP, so that files produced by SPLATCHEcould be also analyzed by these softwares after file conversion.

4.4.2 Nexus files

Three other types of file produced by Friction are compatible with the NEXUS file format: two files with the "*.*trees*" extension are automatically produced and list all the simulated trees, with branch lengths expressed either *i*) in units of generations scaled by the population size (*N*), and therefore representing the true coalescent history of the sample of genes, or *ii*) in units of average number of substitutions per site, and therefore representing the realized mutational tree. These two files could be visualized with the software TREEVIEW (Page 1996). For each simulation, a file with "**paup*" extension could be generated. This file lists all the simulated genes together with their true genealogical structure. This file can be analyzed with David Swofford's PAUP* software (1999). A PAUP batch file, with extension "*.*bat*" is also generated.

4.4.3 Coalescence distribution files

A bitmap representing the spatial distribution of the coalescent events for all the simulations joined is automatically created with the *"*_TotNumCoal.bmp"* termination. This maps can also be visualized by means of the button "Draw Coalescence" (15 on Figure 4.1) on the interface. In setting the coalescence checkbox (16 on Figure 4.1), similar bitmaps of the spatial distribution of coalescent events are generated for every simulations (with the *"*_NumCoal.bmp"* termination). The times for each coalescent event and each simulation are listed on a file with *"*.coal"* extension. Those times are given on generation units, with the bigger number corresponding to the end time of the simulation.

4.4.4 Coalescent trees files

In settings the checkbox "coalescent trees" (16 on Figure 4.1), it is possible to generate for each simulation a bitmap representing the genealogical links between each node of the coalescence tree, laid out spatially. SPLATCHE is the first program which allows an the spatial representation of the coalescent trees. Those files are terminated with *"*_CoalTree_*.bmp"*.

4.4.5 MRCA files

SPLATCHE gives information on the localization and timing of the Most Recent Common Ancestor (MRCA) of the totality of genes sampled, thus on those of the various samples. A file with the termination "*_MRCADensity.bmp*" is automatically generated and is a bitmap of the spatial distribution of MRCA for all the simulations joined. These maps can also be visualized by means of the button "Draw MRCA" (15 on Figure 4.1) on the interface. Similar bitmaps, with the "*_MRCAPopDensity*.bmp*" termination, are generated for each sample. The Time for the Most Recent Common Ancestor (TMRCA) for the whole tree and for each sample are also listed in a file with the *"*.tmrca"* extension. The TMRCA are given on generation units, with the bigger number corresponding to the end time of the simulation.

4.4.6 Other files

The "*gen*" file summarize statistics about the data, such as the mean coalescence times, the mean number of pairwise differences within and among demes and the mean length of the trees.

5 Acknowledgements

We are grateful to Stefan Schneider and Pierre Berthier for their computing assistance. The development of the SPLATCHE program was possible through a Swiss NSF grant n° 31-054059.98.

6 Download sites

SPLATCHE: http://cmpg.unibe.ch/software/splatche/
SIMCOAL: http://cmpg.unibe.ch/software/simcoal/
ARLEQUIN:http://cmpg.unibe.ch/software/arlequin/
TREEVIEW: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
PAUP: http://paup.csit.fsu.edu/

7 References

Currat, M. (in prep).Thèse Département d'Anthropologie et d'Ecologie. Université de Genève.

Currat, M., Ray, N. & Excoffier, L. (2004). SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Molecular Ecology Notes 4(1): 139-142.

Donnelly, P. & Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29: 401-421.

Ewens, W.J. (1990). Population Genetics Theory **-** The Past and the Future. In Kluwer Academic Publishers Mathematical and Statistical developments of Evolutionary Theory: 177-227 S. Lessar. Dordrecht.

Excoffier, L., Novembre, J. & Schneider, S. (2000). SIMCOAL: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J. Heredity 91: 506-510.

Hudson, R. (1990). Gene genealogies and the coalescent process 7?, Oxford University Pressoxford.

Jin, L. & Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7: 82-102.

Jukes, T. & Cantor, C. (1969). Evolution of protein molecules. In Academic press Mamalian Protein Metabolism: 21-132 H.N. Munro. New York.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 111-120.

Kimura, M. & Weiss, W.H. (1964). The stepping stone model of genetic structure and the decrease of genetic correlation with distance. Genetics 49: 561-576.

Kingman, J.F.C. (1982). The coalescent. Stoch. Proc. Appl. 13: 235-248.

Kingman, J.F.C. (1982). On the genealogy of large populations. J. Appl. Proba. 19A: 27-43.

Page, R.D.M. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12: 357-358.

Ray, N. (2003). Modélisation de la démographie des populations humaines préhistoriques à l'aide de données environnementales et génétiques.Thèse Départment d'Anthropologie. Université de Genève.

Ray, N., Currat, M. & Excoffier, L. (2003). Intra-deme molecular diversity in spatially expanding populations. Molecular Biology and Evolution 20(1): 76-86.

Schneider, S., Roessli, D. & Excoffier, L. (2000). Arlequin: a software for population genetics data analysis. User manual ver 2.000. Geneva, Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

ANNEXE 2 Aspects techniques du programme SPLATCHE

Cette annexe est consacrée à la description technique du programme SPLATCHE. Nous y décrivons dans un premier temps les deux principales parties qui composent SPLATCHE : le module démographique (Annexe 2.1) et le module génétique (Annexe 2.2) et dans un second temps la structure du programme lui-même, d'un point de vue informatique (Annexe 2.3).

Annexe 2.1 Module démographique

Comme nous l'avons déjà mentionné dans l'introduction de ce chapitre, l'implémentation de la partie démographique de SPLATCHE est principalement le fait de Nicolas Ray, et a déjà largement été décrite dans sa thèse (Ray 2003). Nous ne reviendrons donc ici que brièvement sur cet aspect de SPLATCHE.

Automate cellulaire

L'aire géographique virtuelle dans laquelle ont lieu les simulations peut être représentée comme une grille régulière (grid) composée de cellules de forme et de taille identiques. Cette structure est communément appelée automate cellulaire (Ray 2003). Chaque cellule représente une petite unité de surface et est associée à plusieurs caractéristiques environnementales, notamment sa position géographique (latitude et longitude). Nous différencierons le terme "**cellule**", qui contient les caractéristiques physiques d'une aire donnée, du terme "**dème**"[1] qui contient les caractéristiques de la population qui peuple la cellule. Cette nomenclature se retrouve dans l'implémentation du programme (Annexe 2.3).

Incorporation des données environnementales

L'utilisation de données environnementales se fait par le biais de cartes numériques de format ASCII raster. Ces cartes peuvent être générées à l'aide de la plupart des logiciels GIS[2]. Trois types d'informations environnementales peuvent être utilisées dans SPLATCHE : la végétation, l'hypsométrie (relief) et l'hydrographie (fleuves, mers et océans). Ces trois types d'information sont connus pour chacune des cellules et peuvent être différents selon les cellules (monde hétérogène). A chaque dème sont associées deux variables, K et F qui résument l'effet des caractéristiques environnementales sur la population. La capacité de soutien K représente le nombre maximum d'individus qui peuvent coexister dans un dème en fonction des ressources de ce dernier. La "friction" F représente la difficulté de mouvement à l'intérieur d'une cellule. D'amples explications sur

[1] Un dème est une sous-population homogène à l'intérieure de laquelle le choix des partenaires se fait de manière aléatoire (Gilmour et Gregor 1939).

[2] Par exemple avec le logiciel ARCVIEW (ESRI 1998).

la façon dont les paramètres *F* et *K* peuvent être calculés, en fonction des données environnementales, sont données dans Ray (2003 : chapitre 3).

Unité de temps

Les simulations se font par une succession d'itérations temporelles qui correspondent à des générations discrètes[1]. Lors de chaque génération, quatre étapes (Figure 9.1) ont lieu successivement dans chacun des dèmes qui composent le monde virtuel :

1. Régulation de la densité N_t , selon une croissance logistique;
2. Calcul du nombre effectif d'émigrants *E*;
3. Calcul de la direction de migration pour chaque émigrant;
4. Mise à jour de la densité N_{t+1}, au temps $t+1$, en fonction du nombre d'émigrants *E* et d'immigrants *I* provenant des cellules voisines :

$$N_{t+1} = N_t - E + I$$

Evolution de la densité

Le nombre d'individus qui peuplent un dème peut évoluer au cours du temps selon différents modèles démographiques. Ces modèles sont décrits de façon détaillée dans l'ANNEXE 1, ainsi que dans Ray (2003).

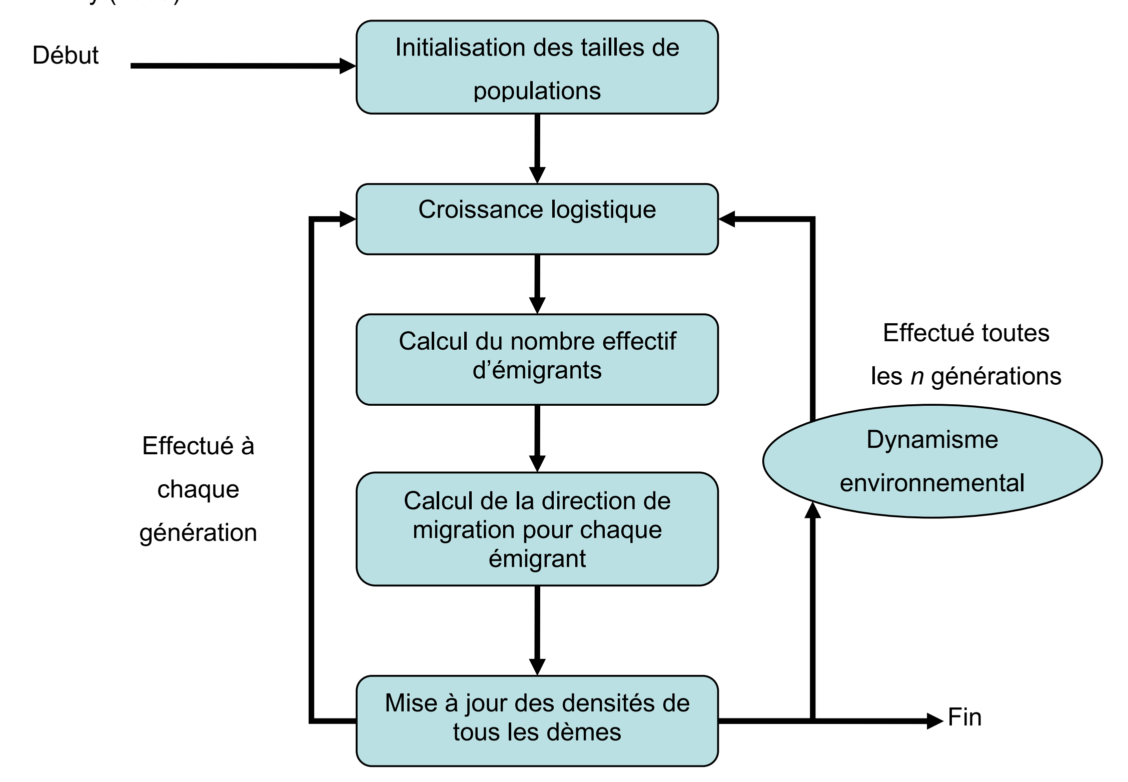


Figure 9.1. Étapes principales d'une simulation démographique.

---

[1] L'hypothèse est faite que chaque nouvelle génération d'individus succède entièrement à la précédente, sans qu'il y ait de superposition entre deux générations successives.

Migrations

Les dèmes sont arrangés selon un modèle appelé "stepping-stone" en deux dimensions (Kimura 1953), ce qui signifie qu'un individu appartenant à un dème donné peut potentiellement migrer dans chacun des quatre dèmes voisins. Ce modèle prend donc en compte la disposition spatiale des sous-populations (Figure 9.2). Le nombre d'émigrants est simplement une proportion *m* de la densité N_t et leur direction est calculée en fonction de la friction (*F*) des cellules voisines (voir la section 4.3.2.1, ainsi que l'Annexe 1 pour les détails).

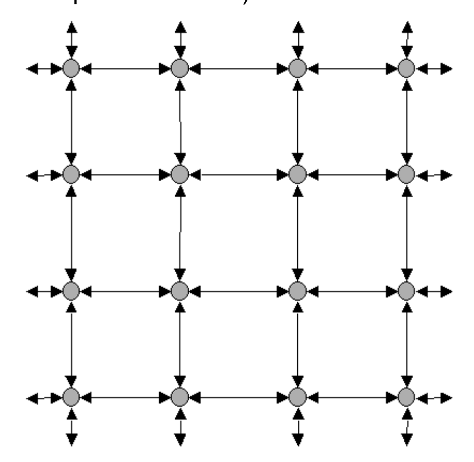


Figure 9.2. Schéma du modèle "stepping-stone 2D".

Base de données démographique

Toutes les étapes démographiques ne sont pas sauvegardées dans la base de données virtuelle. Seules le nombre d'immigrants et la densité finale N_{t+1} sont stockés lors de chaque génération, pour chaque dème. La base de données contient donc *D x G x 5* éléments, où *D* correspond au nombre de dèmes et *G* au nombre de générations simulé. Le nombre 5 représente, pour chaque dème, la densité *N* ainsi que le nombre d'immigrants ($N_I = m_E + m_O + m_N + m_S$) provenant de chacune des 4 directions (Est, Ouest, Nord et Sud). Le nombre d'immigrants est stocké, contrairement au nombre d'émigrants, puisque la phase de coalescence se déroule en remontant le temps. Il est donc nécessaire de connaître le nombre de gènes étant arrivés dans un dème, ainsi que leur dème d'origine (voir aussi l'Annexe 2.1).

Dynamique environnementale

La version publique de SPLATCHE permet uniquement la simulation d'un monde statique, pour lequel les variables environnementales ne fluctuent pas au cours du temps. En revanche, la version en développement du logiciel ("FRICTION") permet de modifier ces variables toutes les *n* générations (Figure 9.1). Cette implémentation permet de modéliser la dynamique de l'environnement sous l'effet des variations du climat. Les changements du niveau des mers, ainsi que les changements de végétation lors des périodes de réchauffement ou de refroidissement peuvent ainsi être pris en compte. Nous n'aborderons pas de façon plus approfondie cet aspect de

"FRICTION" ici, puisque nous ne l'utilisons pas directement dans ce travail et qu'il a été décrit de façon détaillée dans Ray (2003).

Annexe 2.2 Module génétique

Le module génétique ne peut pas être utilisé indépendamment du module démographique puisqu'il utilise la base de données générée par ce dernier comme paramètre d'entrée. Le module génétique est dérivé du logiciel SIMCOAL (Excoffier *et al.* 2000), dont les matrices de densité et de migration ont été remplacées par la base de données générée par le module démographique de SPLATCHE. Ce dernier permet de simuler des données génétiques pour un éventail de scénarios démographiques bien plus large que SIMCOAL. L'intégration des routines de SIMCOAL dans SPLATCHE a été longue et délicate puisqu'il a été nécessaire de les adapter à des modèles démographiques beaucoup plus complexes. L'une des difficultés principales a été de traduire les données démographiques en nombres entiers, nécessaires au fonctionnement des fonctions de coalescence, tout en gardant une cohérence stricte entre nombres de migrants et densités. L'utilisation de nombres entiers a également nécessité une implémentation spécifique, afin que la précision numérique des modèles démographiques reste satisfaisante.

Une simulation génétique se divise en deux phases. Lors de la première, la généalogie d'un nombre arbitraire de gènes échantillonnés dans le monde virtuel est reconstruite à l'aide d'un algorithme dérivé de la théorie de la coalescence. Puis, lors de la seconde phase, des mutations sont superposées aléatoirement sur les branches de l'arbre généalogique obtenu (ou arbre de coalescence) afin de générer la diversité moléculaire des gènes échantillonnés.

Annexe 2.2.1 Processus de coalescence

La première phase d'une simulation génétique utilise la théorie de la coalescence, qui a été formalisée en 1982 par Kingman (1982a; 1982b), puis développée ultérieurement (Ewens 1990 ; Hudson 1990 ; Donnelly et Tavaré 1995 ; Nordborg 2001). Cette approche permet de reconstruire, en remontant dans le temps, la généalogie d'une série de gènes échantillonnés au présent jusqu'à leur ancêtre commun le plus récent, leur MRCA (<u>M</u>ost <u>R</u>ecent <u>C</u>ommon <u>A</u>ncestor, Figure 9.3).

Les "**lignages**" des gènes, ou leur généalogie, est reconstruite jusqu'au moment où deux d'entre eux possèdent un ancêtre commun. Il s'agit d'un "**événement de coalescence**". Cette coalescence a lieu dans un individu qui a transmis obligatoirement au moins deux copies d'un même gène à des descendants distincts. Après chaque coalescence, le nombre de lignages diminue d'une unité. Il est donc possible de continuer ce processus jusqu'au moment où il n'existe plus qu'un seul lignage, le MRCA. Le temps qui s'écoule du présent jusqu'au MRCA est appelé TMRCA (<u>T</u>ime to the <u>M</u>ost <u>R</u>ecent <u>C</u>ommon <u>A</u>ncestor). Lorsque *n* gènes sont échantillonnés, *n*-1 événements de coalescence ont lieu. A la fin du processus, un arbre de coalescence est constitué, dont la racine est le MRCA et les branches terminales sont les gènes échantillonnés.

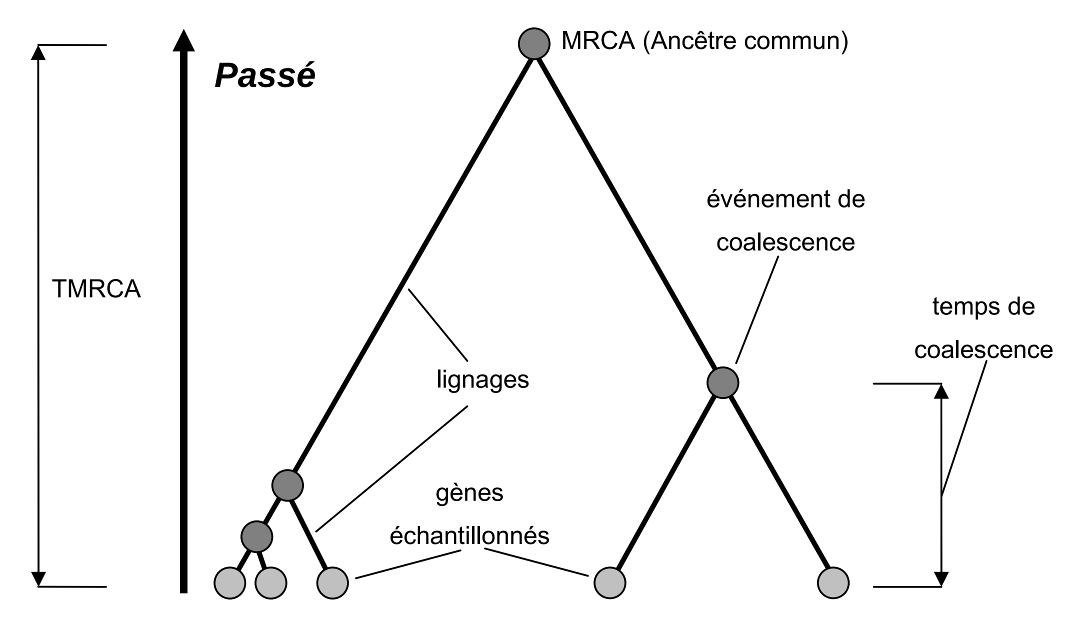


Figure 9.3. Exemple d'un arbre de coalescence de cinq gènes échantillonnés.

L'arbre de coalescence représente donc la généalogie des gènes, qu'il ne faut pas confondre avec une généalogie d'individus, même si les deux sont intimement liées (voir par exemple Excoffier 1997 ; Hurles et Jobling 2001 ; Nichols 2001). Chaque copie d'un gène possède une généalogie spécifique, dont aucune ne ressemble à celle des individus, comme l'illustre la Figure 9.4. Dans cette figure, les deux exemplaires du gène représenté par un rond noir, que l'on trouve à la génération 3, ont par exemple un ancêtre commun à la génération 1, alors que le gène représenté par un rond gris, ne laisse aucun descendant à la génération 3.

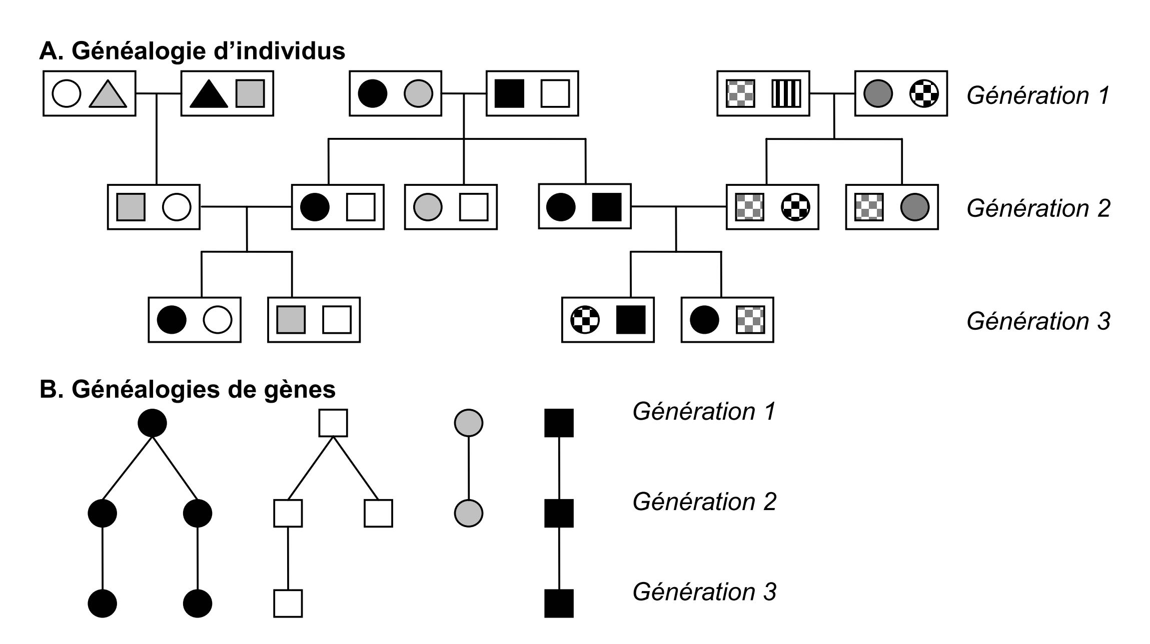


Figure 9.4. Schéma d'une généalogie d'individus diploïdes, pendant 3 générations (A), ainsi que les généalogies de 4 copies différentes d'un gène (B).

La généalogie d'un gène sélectivement neutre[1] dépend uniquement de l'histoire démographique de la population dans laquelle il se trouve et il est complètement indépendant du processus de mutation. L'étude de nombreuses généalogies de gènes neutres permet donc de faire des inférences sur l'histoire démographique de la population dans laquelle ils sont échantillonnés. L'intérêt principal de la théorie de la coalescence réside donc dans l'étude de la démographie d'une population à partir d'un échantillon de ses gènes. Il s'agit de l'avantage principal apporté par l'implémentation de l'approche par coalescence, par rapport à une approche "forward" (Livingstone 1989 ; Kaplan *et al.* 1991 ; Currat *et al.* 2002 ; Edmonds *et al.* 2004), puisqu'elle ne rend pas nécessaire la simulation de la totalité des gènes d'une population, mais seulement de ceux qui sont échantillonnés, ainsi que de leurs ancêtres. Le gain en temps de calcul et en espace mémoire est donc très important.

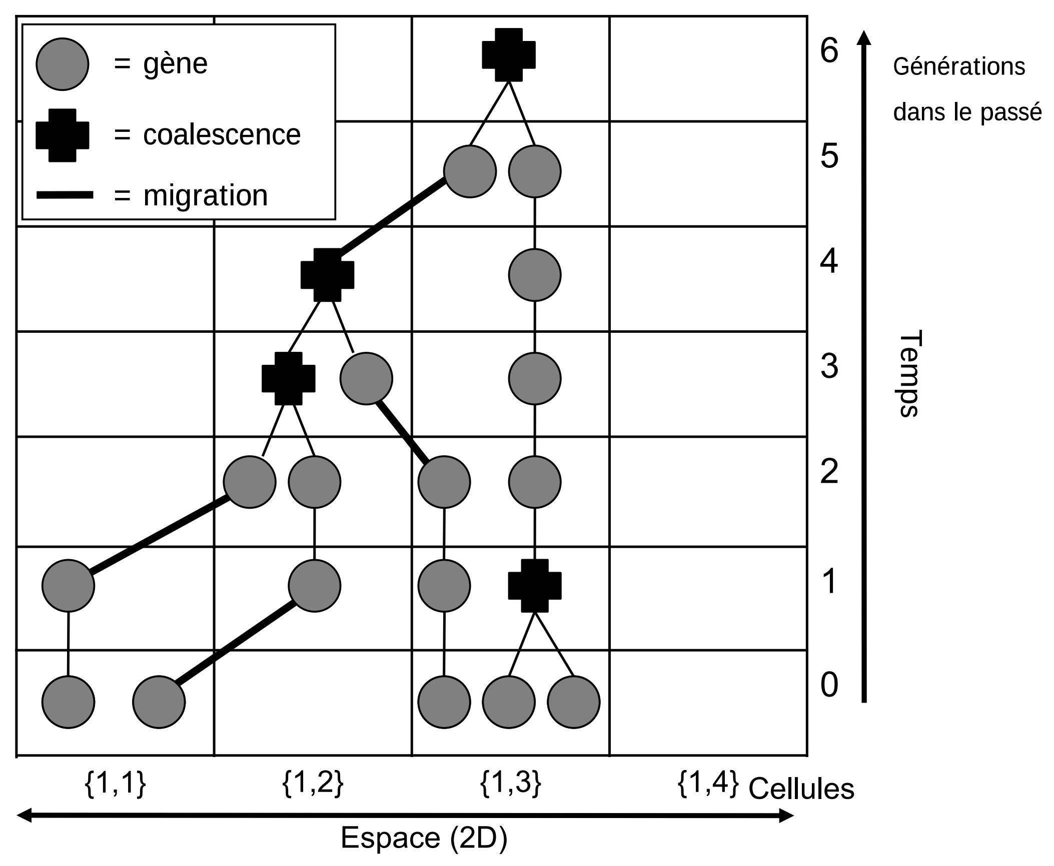


Figure 9.5. Schématisation du processus de coalescence aboutissant à un arbre de coalescence des gènes échantillonnés, tel qu'il est implémenté dans SPLATCHE. Dans cette figure, 2 gènes sont, par exemple, échantillonnés dans le dème {1,1} mais leur coalescence a pourtant lieu dans un dème différent {1,2} à la génération -3. Ceci souligne la diffusion spatiale des lignages au cours du temps.

L'implémentation du processus de coalescence dans SPLATCHE se déroule de la façon suivante : lors de chaque génération, chacun des lignages simulés peut subir deux types d'événements : 1) une coalescence ou 2) une migration (Figure 9.5).

[1] Un gène neutre est considéré comme n'étant pas sujet à la sélection.

1) Evénement de coalescence :

Pour qu'une coalescence soit possible, il faut qu'au moins deux lignages se trouvent dans le même dème au même moment. Comme un dème avec *N* individus diploïdes contient 2*N* copies d'un gène donné, la probabilité d'avoir un ancêtre commun à la génération précédente est égale à 1/(2*N*) pour chaque paire de gènes présents dans ce dème. Si la population est haploïde, cette probabilité est égale à 1/*N*. La probabilité de coalescence P_c est donc inversement proportionnelle à la taille N_t de la population et dépend uniquement de cette dernière. Dans SPLATCHE, la densité N_t est tirée de la base de données virtuelle générée pendant la phase démographique. Si *n* lignages se trouvent dans un dème, alors il existe $n(n-1)/2$ paires de lignages possibles. La probabilité d'avoir un événement de coalescence pendant cette génération devient $n(n-1)/(4N)$ lorsque la population est diploïde et $n(n-1)/(2N)$ lorsqu'elle est haploïde. Une seule coalescence par dème et par génération est possible dans SPLATCHE. Lorsqu'une coalescence se produit, alors le nombre total de lignages présents dans le monde virtuel est décrémenté d'une unité.

2) Migration :

Chaque lignage appartenant au dème *i* a une probabilité m_{ij} de migrer vers le dème voisin *j* au temps *t*. Cette probabilité se calcule comme

$$m_{ij} = \frac{I_{ji}}{N_i} \qquad \text{(Eq. 9.1)}$$

où I_{ji} est le nombre d'immigrants qui sont arrivés dans le dème *i* depuis le dème *j* pendant la phase démographique et N_t est la densité de population du dème à la génération *t*. La migration d'un lignage de *i* vers *j*, en remontant dans le temps, correspond au fait que ce lignage a été apporté dans *i* par un immigrant venu de *j*. m_{ij} est calculée pour tous les voisins d'un dème (de un à quatre). Le nombre d'immigrants I_{ji} provenant de chacun des dèmes voisins est tiré de la base de données générée pendant la phase démographique (somme des migrants venant des 4 points cardinaux, voir page 175).

Annexe 2.2.2 Génération de la diversité génétique

La seconde phase d'une simulation génétique consiste en la génération de la variabilité des échantillons. En effet, pendant la première phase, tous les gènes échantillonnés sont considérés comme identiques d'un point de vue génétique. Seul un identificateur numérique permet de les reconnaître les uns des autres, et ainsi de reconstruire leurs liens généalogiques. Dans la seconde phase, un haplotype[1] est défini aléatoirement, constitué d'un type de données génétiques choisi par l'utilisateur (séquence d'ADN, RFLP ou microsatellite). Cet haplotype est introduit à la racine de l'arbre de coalescence. Puis, des mutations sont simulées le long des branches de l'arbre, en fonction d'une distribution de Poisson centrée sur μt, où μ est le taux de mutation et *t* la longueur

[1] Un haplotype est une combinaison unique d'un marqueur génétique présent sur un chromosome (Hartl et Clark 1997 : p. 57).

d'une branche en générations (Figure 9.6). Plus une branche est longue, et plus elle aura de chance de porter une mutation. Cela signifie qu'un gène échantillonné qui se trouve au bout d'une très longue branche terminale aura une grande probabilité d'avoir des mutations propres et d'être très différencié des autres gènes. Le taux de mutation est choisi en fonction du nombre de locus liés qui constituent l'haplotype.

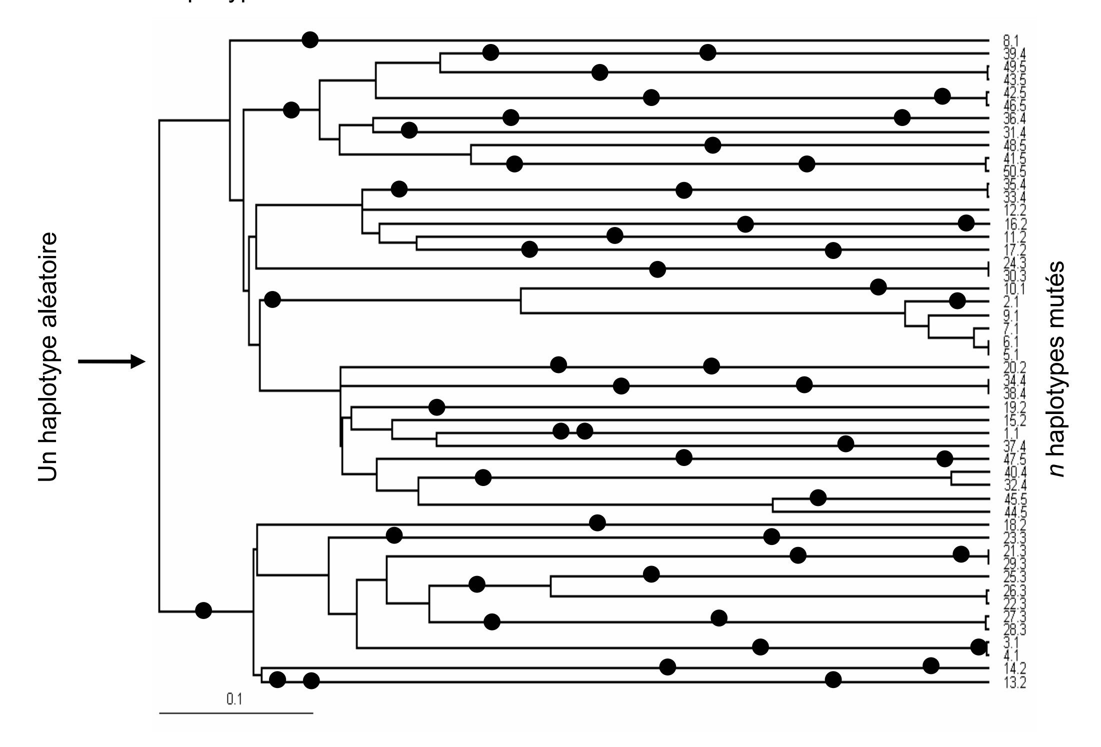


Figure 9.6. Exemple d'arbre de coalescence. Les points représentent des mutations aléatoires le long des branches de l'arbre, aboutissant à des séquences mutées.

SPLATCHE offre donc la possibilité de générer des données moléculaires, lesquelles dépendent de la topologie d'un arbre de coalescence : topologie elle-même influencée par la démographie de la population. La Figure 9.7 illustre la forme caractéristique des arbres de coalescence dans le cas de deux démographies de populations différentes (stationnaire ou ayant passé par une croissance démographique), ainsi que les données moléculaires obtenues, sous la forme de distributions "mismatch" et de distributions de fréquences alléliques.

La Figure 9.8 montre que des réplications indépendantes d'arbres de coalescence selon un scénario démographique donné sont sujettes à une grande variance, puisque la stochasticité du processus est grande. Cependant, la topologie générale des arbres reste tout de même reconnaissable si les scénarios démographiques sont bien différenciés. Une expansion démographique donne, le plus souvent, des arbres en forme de peigne, avec de très longues branches terminales (Figure 9.8B). En revanche, un population stationnaire présente une très grande variabilité dans la topologie des arbres, avec une alternance de courtes et de longues branches terminales (Figure 9.8A). L'augmentation du nombre de simulations génétiques permet de tenir compte de la stochasticité des processus démographiques et génétiques.

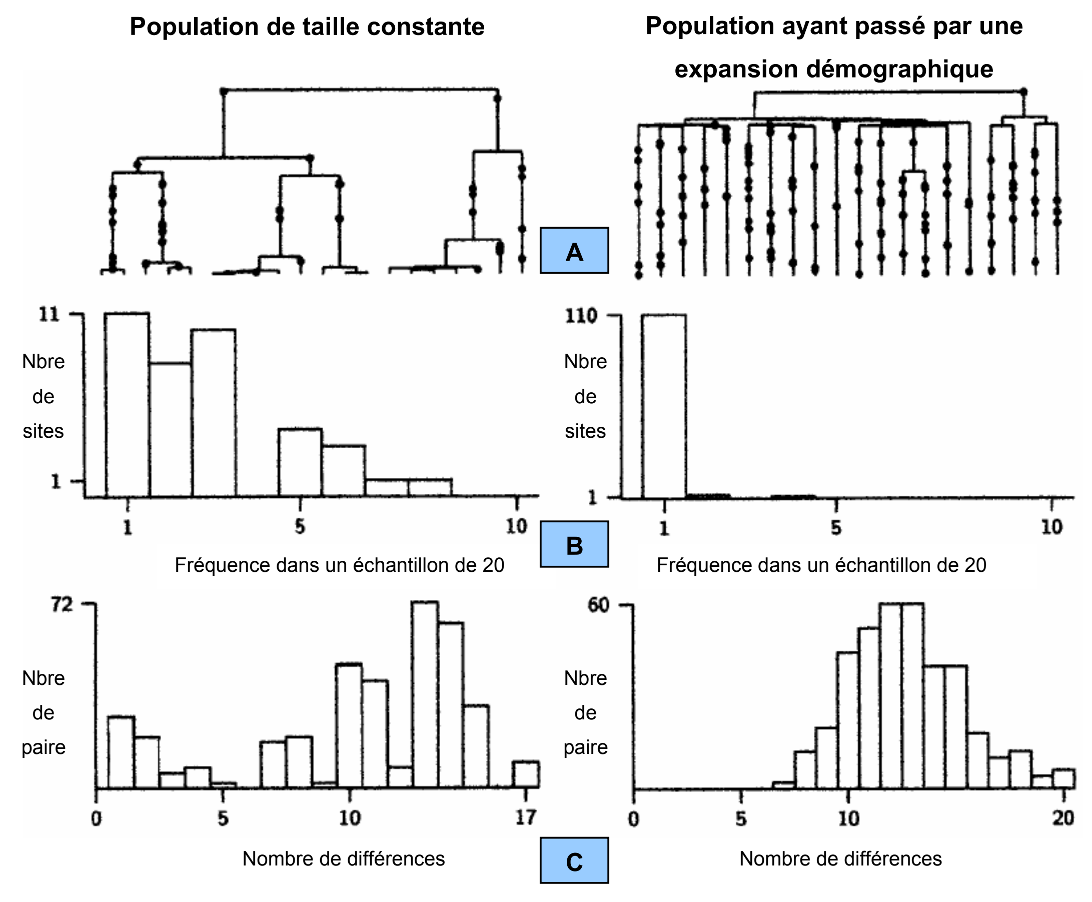


Figure 9.7. Relation entre généalogies de gènes et données moléculaires pour une population dont la taille est restée constante au cours du temps (colonne de gauche), et pour une population ayant passé par une expansion démographique (colonne de droite). A : arbres de coalescence ; B : spectres de fréquences alléliques ("Allele Frequency Spectrum") ; C : distribution "mismatch ". Figure modifiée, à partir de Harpending *et al.* 1998.

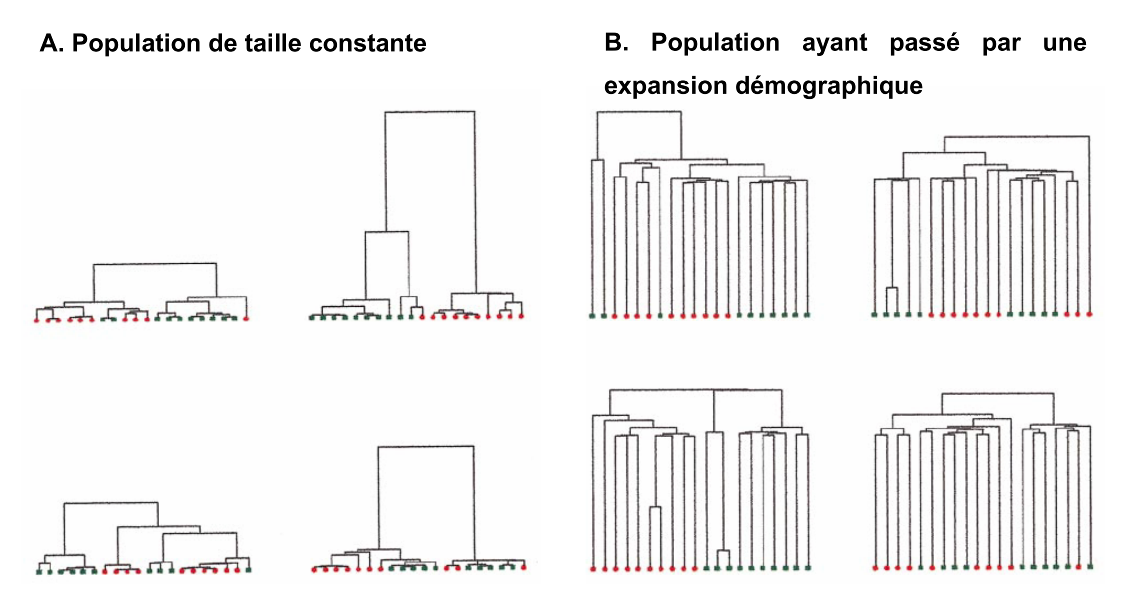


Figure 9.8. Exemples de 4 arbres de coalescence indépendants, obtenus par la simulation d'une population stationnaire (A) et par celle d'une population ayant passé pat une expansion démographique importante (B). Figure modifiée à partir de Harpending *et al.* 1998.

Annexe 2.2.3 Génération de SNPs

Nous avons modifié la version de SPLATCHE mise à la disposition du public et présentée dans le chapitre 2, afin de permettre la simulation de données de type SNP dont l'utilisation est de plus en plus fréquente. Afin d'économiser un temps de calcul important, l'implémentation permettant la création des SNPs est différente de celle qui permet la création des autres types de données (Annexe 2.2.2). En effet, les SNPs se différencient des autres données par le fait que tous les locus sont polymorphes. Afin d'éviter de simuler un grand nombre de locus et de ne garder que ceux qui sont polymorphes, nous avons opté pour une autre stratégie. Une fois l'arbre de coalescence construit selon la description faite dans l'Annexe 2.2.1, la longueur totale *L* (en générations) des branches qui composent l'arbre est comptée. Un nombre *x*, situé dans l'intervalle]0 ; *L*] est ensuite tiré aléatoirement. En fonction de *x*, une branche de l'arbre est choisie et une mutation y est superposée. Plus une branche est longue et plus la probabilité qu'une mutation y apparaisse est grande. Les *p* gènes échantillonnés, issus de lignages descendant de celui sur lequel se trouve la mutation, présenteront ainsi l'état muté du SNP. La fréquence de ce SNP dans la population sera donc égale à *p*/*n*, où *n* est le nombre total de gènes échantillonnés. Cette implémentation permet l'économie de la simulation de tous les locus monomorphes.

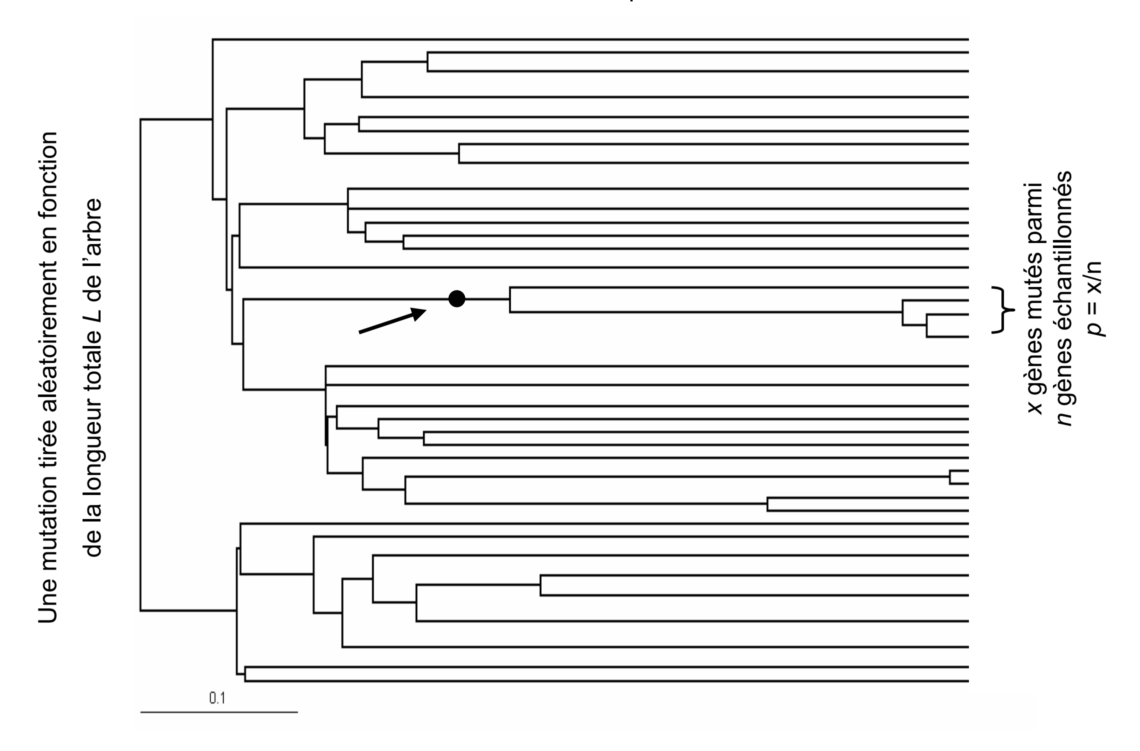


Figure 9.9. Exemple du choix aléatoire d'un SNP (rond noir), en fonction de la longueur totale de l'arbre de coalescence. Les *x* gènes échantillonnés descendant du lignage sur lequel la mutation est apparue présentent l'état muté du SNP.

Cette implémentation permet également de simuler des SNPs sujets à un "**biais de recrutement**" ("ascertainment bias" en anglais). Ce biais de recrutement signifie que les mutations typées dans les échantillons ne sont pas représentatives de celles de la population puisque les mutations les plus rares dans la population sont sous-représentées dans les échantillons (Rogers et

Jorde 1996). Ce biais de recrutement peut être dû à trois phénomènes. Tout d'abord, la probabilité d'observer une mutation dépend de sa fréquence dans la population. De plus, les SNPs dont la représentation est inférieure à un certain pourcentage (5%,10%) sont souvent écartés des analyses, car jugés non informatifs (Casalotti *et al.* 1999 ; Underhill *et al.* 2000 ; Akey *et al.* 2002). Finalement, la détection d'un SNP peut être faite dans une sous-population particulière et celui-ci peut être monomorphe dans une autre sous-population. Par exemple, un polymorphisme détecté dans les populations européennes ne sera peut-être pas variable dans les populations africaines. Il apporterait donc une information biaisée s'il était utilisé pour étudier la variabilité des populations africaines.

La version modifiée de SPLATCHE permet de choisir un fréquence minimale P_{min} d'un SNP dans l'ensemble des échantillons ou dans au moins un échantillon[1]. De cette manière, les SNPs sont tirés aléatoirement, mais seuls ceux dont $p > P_{min}$ sont gardés. Le biais de recrutement peut modifier de façon importante l'interprétation des données (voir sections 3.3 et 6.2.1). Il est donc nécessaire de le simuler afin de connaître son influence sur la signature génétique laissée par des événements démographiques.

Annexe 2.3 Implémentation

Nous ne rentrerons pas ici dans les détails de l'implémentation du programme SPLATCHE, mais nous décrirons simplement les points principaux de sa structure. Le programme SPLATCHE a été développé en C++, qui est un langage orienté objet. Par conséquent, le code est constitué de différentes classes qui permettent la création d'objets lors d'une simulation.

Annexe 2.3.1 Principales classes

- World :

La classe *World* est à la base d'une simulation puisqu'elle représente le monde virtuel lui-même. C'est à l'intérieur de *World* que se trouvent toutes les caractéristiques de cette aire virtuelle : la taille des cellules ainsi que leur nombre; la liste des capacités de soutien (*K*) et des coefficients de friction (*F*), en fonction des types d'environnement. *World* dispose d'une matrice de pointeur vers les cellules qui constituent le monde virtuel (*Worldmatrix*), ainsi qu'un pointeur vers la base de données démographique (*S_DB*).

- Cell :

La classe *Cell* représente une cellule appartenant au monde virtuel. Elle possède donc tous les attributs physiques d'une aire géographique donnée, telles que ses coordonnées (latitude et longitude), son type de végétation, ainsi que le fait qu'elle soit côtière ou traversée par une rivière. De plus, *Cell* possède un membre *State* qui contient l'état courant de la cellule pendant le

[1] P_{min} est égal à la fréquence de l'allèle mineur - le moins fréquent des deux états du SNP - dans la population.

déroulement de la simulation démographique. La classe *State* contient la densité courante de la cellule, ainsi que les variables *K* et *F*, qui peuvent varier au cours du temps si le dynamisme environnemental est utilisé. En plus de ses coordonnées géographiques, *Cell* est défini par un index numérique (*CellIndex*), qui permet d'y accéder de façon plus rapide, notamment à partir du module génétique. C'est également la classe *Cell* qui contient l'implémentation des modèles démographiques et les différentes variables qui y sont liées.

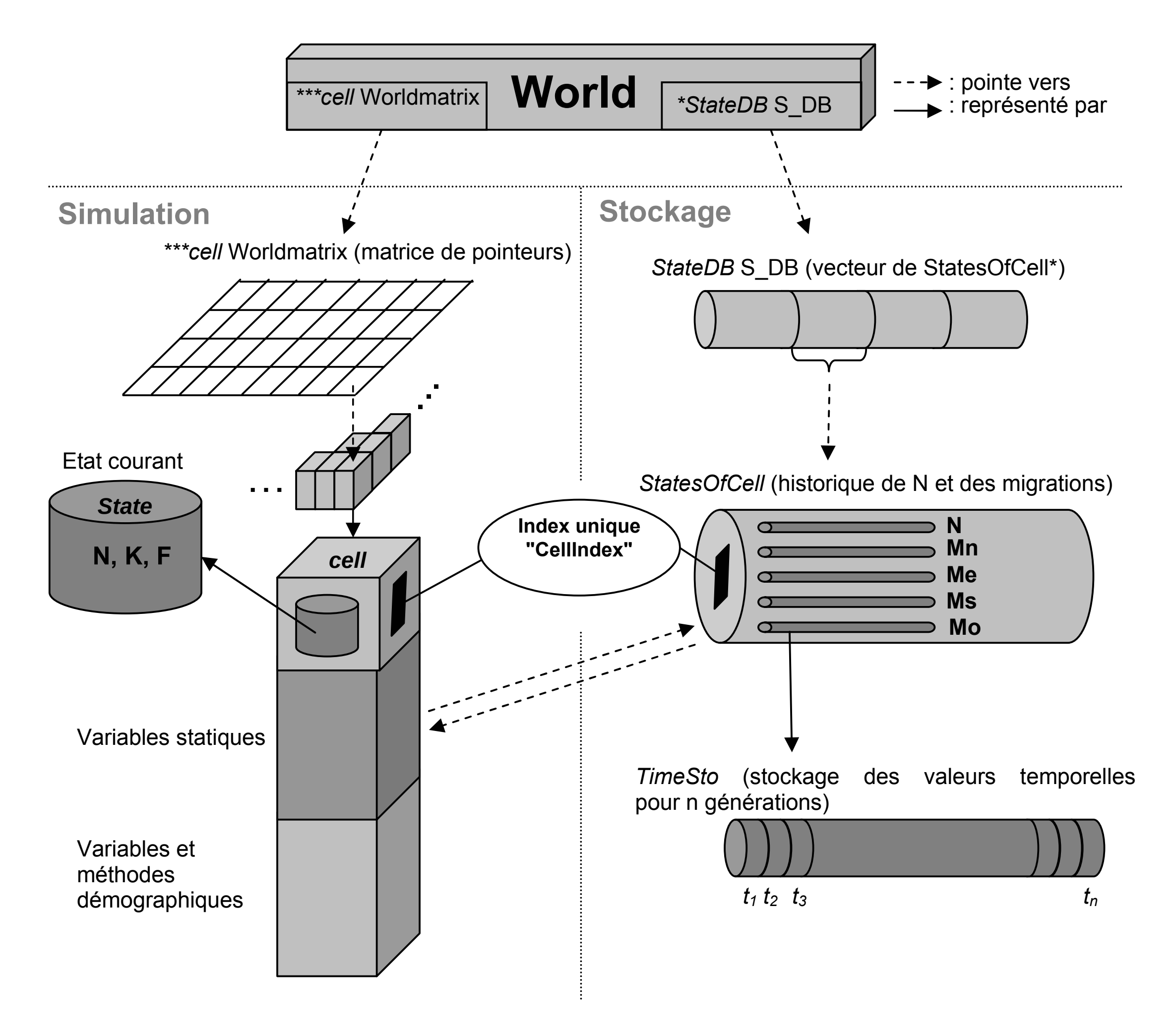


Figure 9.10. Schéma des objets principaux et de leurs relations pour le module démographique de SPLATCHE. Les noms des classes des objets sont en italiques. Voir le texte pour les explications.

- StateDB :

StateDB est la classe qui contient la base de données démographiques. Elle possède, notamment, un vecteur de pointeurs de dimension *n*, où *n* est égal au nombre de *Cell* simulées. Ces pointeurs sont dirigés vers des objets *StatesOfCell*, dont chacun contient l'historique d'une cellule donnée. La position d'un *StatesOfCell* dans le vecteur est égale à l'index numérique *CellIndex* de l'objet *Cell* correspondant. Chaque *StatesOfCell* contient 5 objets *TimeSto*, qui sont des vecteurs de dimension *t*, où *t* est égal au nombre de générations simulées. Quatre *TimeSto* stockent les

immigrants provenant de chacune des cellules voisines, alors que le 5[ème] stocke la densité N_t de la cellule.

- TDemeCollection :

La classe *TDemeCollection* est l'équivalent de la classe *World* pour tout ce qui concerne les simulations génétiques. Cette classe contient toutes les caractéristiques nécessaires à une simulation génétique, notamment un vecteur de pointeurs, de dimension *n*, où *n* est égal au nombre d'objets *Cell* simulés. Chacun des pointeurs appartenant à ce vecteur est dirigé vers un objet *TDeme*, dont la position correspond à l'index numérique *CellIndex* de l'objet *Cell* correspondant. L'arbre de coalescence virtuel *Ttree* est également un membre de *TDemeCollection. Ttree* est un vecteur contenant 2*n*-1 *TNode*, où *n* est le nombre de gènes échantillonnés dans le monde virtuel. Chaque *TNode* représente un nœud de l'arbre de coalescence et possède deux pointeurs "descendants" et un pointeur "ancêtre", qui représentent les liens généalogiques qui unissent les différents nœuds de l'arbre de coalescence.

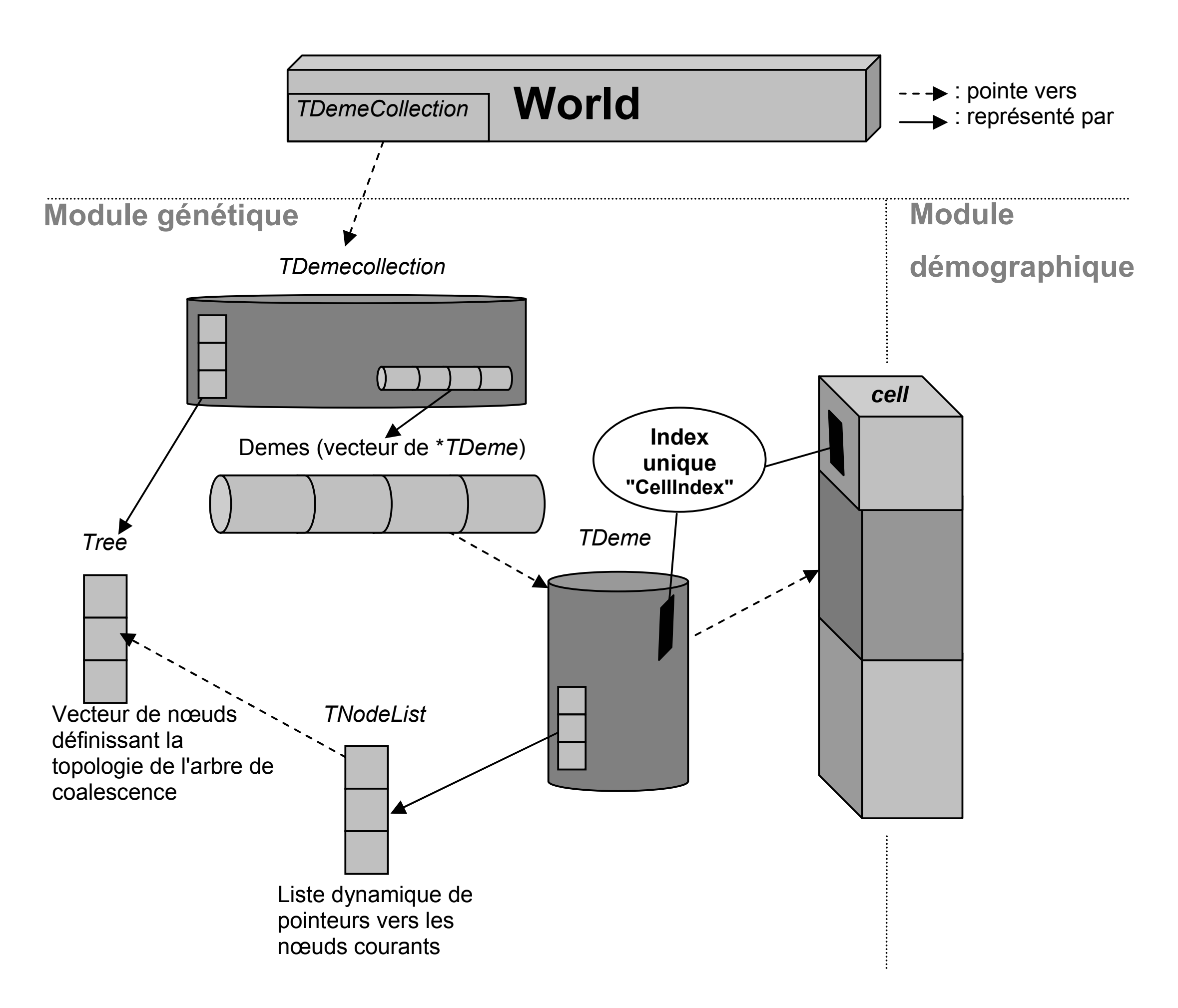


Figure 9.11. Schéma des objets principaux et de leurs relations pour le module génétique de SPLATCHE. Les noms des classes des objets sont en italiques. Voir le texte pour les explications.

- *TDeme :*

La classe *TDeme* est l'équivalent génétique de la classe *Cell*. En effet, si *Cell* représente les caractéristiques physiques d'une aire géographique donnée, *TDeme* représente la population qui peuple cette aire. Par conséquent, il existe un objet *TDeme* par objet *Cell* et des pointeurs permettent d'accéder directement de l'un à l'autre. *TDeme* contient, notamment, une liste des lignages qui se trouvent dans la cellule au temps courant, pendant la simulation génétique (*TNodeList*). *TNodeList* est une liste dynamique de pointeurs vers les *TNode* appartenant à *Ttree*. Cette liste est mise à jour lors de chaque génération, en fonction du mouvement des lignages au cours du temps.

SPLATCHE est bien évidemment constitué de nombreuses autres classes et fonctions, que nous ne décrirons pas ici. Il nous semble cependant important de mettre en avant la complexité de ce logiciel, dont la réalisation a demandé plus de trois ans de travail à un groupe de 3 personnes. Il a fallu une coordination importante afin d'optimiser l'efficacité du logiciel, en fonction des contraintes informatiques, mais également de celles imposées par les modèles démographiques et génétiques. Comme nous l'avons déjà signalé, il existe une version évolutive de SPLATCHE ("FRICTION"), pour laquelle des fonctionnalités sont ajoutées régulièrement.

ANNEXE 3 Visualisation de la coalescence

Une des particularités de SPLATCHE est la possibilité de représenter graphiquement les arbres de coalescence. C'est à notre connaissance le seul logiciel qui permette de visualiser les différentes composantes des généalogies de gènes de façon spatiale. Ces sorties graphiques sont très intéressantes d'un point de vue didactique, car elles permettent de bien comprendre le processus de coalescence, en fonction de différents scénarios démographiques simulés. Trois composantes des généalogies peuvent être superposées à la carte de la région où a lieu la simulation :

Annexe 3.1 Arbre de coalescence

Il est possible de visualiser les liens généalogiques entre les différents nœuds de la généalogie, en fonction des endroits où ont eu lieu les coalescences. Il s'agit donc de la représentation spatiale de l'arbre de coalescence (Figure 9.12).

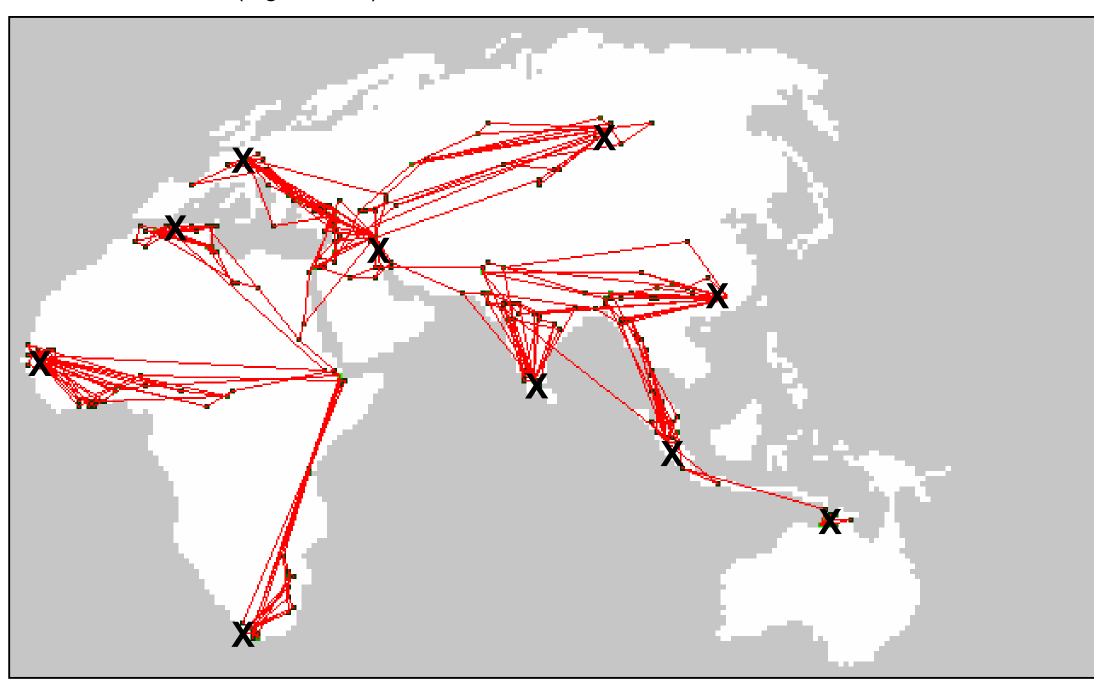


Figure 9.12. Exemple d'un arbre de coalescence obtenu après une simulation dans l'ancien monde. Les croix noires représentent les endroits d'où sont tirés les échantillons.

Annexe 3.2 Distribution des événements de coalescence

SPLATCHE offre la possibilité de visualiser la distribution spatiale des événements de coalescence. Cette représentation permet de mettre en évidence les régions dont les densités de coalescence sont élevées, traduisant ainsi des effets fondateurs importants.

La Figure 9.13 présente la distribution spatiale des coalescences obtenues après 1'000 simulations, en Europe, lorsque 800 gènes sont à chaque fois échantillonnés sur un axe situé entre le Liban et l'Irlande. La simulation consiste en l'expansion démographique et spatiale d'une population originaire du Proche-Orient (flèche blanche) dans une aire homogène pour les facteurs environnementaux. De nombreuses coalescences ont lieu dans le dème source de cette population, comme l'indique le pixel plus clair près de la flèche blanche. La densité de coalescence est particulièrement forte dans les régions où les possibilités de dispersion des lignages sont réduites, comme dans les régions de la Manche et du Détroit du Bosphore (cercles blancs, que nous nommerons "goulets spatiaux"). Il est également possible d'observer des pixels plus clairs le long de l'axe d'échantillonnage. Ces points correspondent aux localisations des échantillons, et représentent les coalescences récentes qui ont lieu dans ces échantillons lors de la phase de disparition (*S1*) ou "scattering phase" (Wakeley 1999, 2001). Les sections 3.2 et 4.5.3 permettent de mieux comprendre la dynamique temporelle des événements de coalescence.

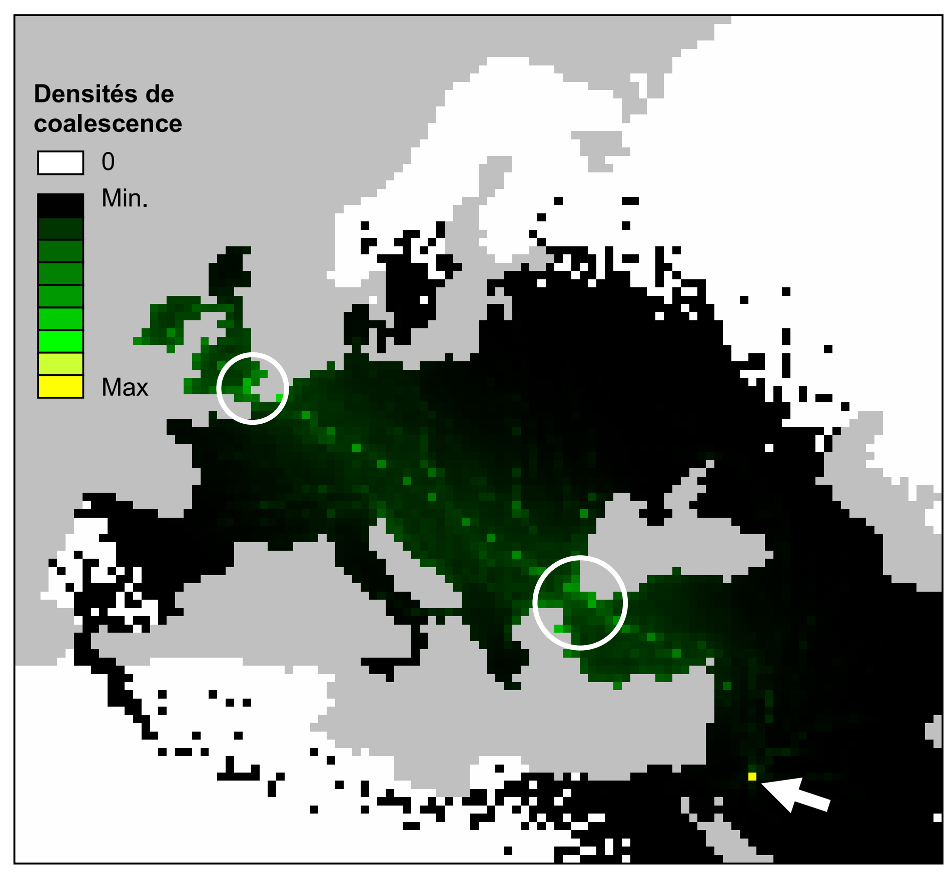


Figure 9.13. Distribution spatiale d'événements de coalescence après l'expansion d'une population originaire du Proche-Orient (flèche blanche). 20 échantillons de 40 gènes sont échantillonnés 1'000 fois le long d'un axe allant du Liban à l'Irlande. Des ponts de terre artificiels sont simulés dans des endroits où le passage d'individus est hautement probable (la Manche, Gibraltar, etc…). Les cercles blancs soulignent les goulets spatiaux dans lesquels la densité de coalescence est importante, ce qui correspond à de forts effets fondateurs lors de la vague de migration.

Annexe 3.3 Distribution des MRCA

SPLATCHE permet également la visualisation de la distribution spatiale des ancêtres communs les plus récents, soit pour les gènes appartenant à un échantillon particulier x ($MRCA_X$), soit pour la totalité des échantillons ($MRCA_T$). La visualisation des $MRCA_X$ (avec x compris entre 1 et 5) pour 5 échantillons de 40 gènes situés à différents endroits de l'Europe permet de souligner l'importance des "goulets spatiaux" (Figure 9.14). En effet, la densité des $MRCA_X$ est très forte après un goulet, dans la direction de la source de la population en expansion (Figure 9.14A-E). Ceci s'explique par le fait que de nombreuses coalescences ont lieu dans la région du goulet (voir ci-dessus), et que, par conséquent, le nombre de lignages qui subsistent après le passage par le goulet est faible. Il suffit que ces derniers lignages coalescent sur le chemin qui les mène à la source de l'expansion, pour que le $MRCA_X$ ait lieu avant d'arriver sur ce lieu d'origine. Par conséquent, plus un échantillonnage est effectué loin de la source d'une expansion et plus la probabilité que le $MRCA_X$ de cet échantillon se trouve le long de l'axe entre le lieu d'échantillonnage et cette source est grande. En revanche, il est extrêmement rare que l'ancêtre commun à tous les échantillons ($MRCA_T$) se trouve sur cet axe, puisque les 1'000 $MRCA_T$ simulés pour l'ensemble des 5 échantillons (Figure 9.14F) se trouvent dans le dème source de la population.

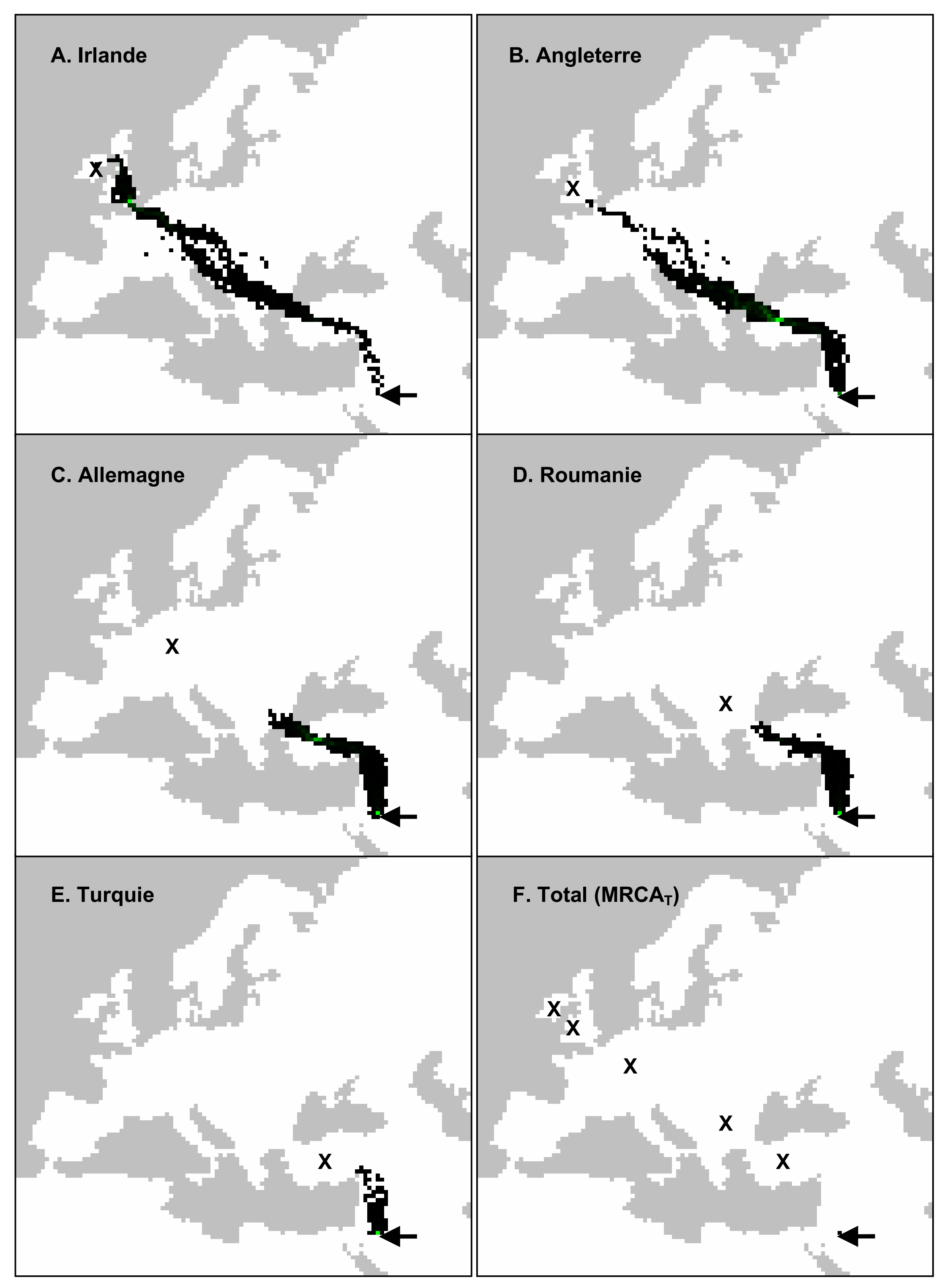


Figure 9.14. A-D: Distribution spatiale de 1'000 MRCA$_X$ (voir texte) simulés pour chacun des 5 échantillons de 40 gènes (croix noires) après la dispersion d'une population depuis le Proche-Orient (flèche noire). F : distribution des 1'000 MRCA$_T$ pour la totalité des 5 échantillons. Les pixels noirs représentent les dèmes dans lesquels ont eu lieu au moins un MRCA, alors que les pixels de terre gris représentent les dèmes dans lesquels la densité de MRCA est la plus élevée.

ANNEXE 4 Modifications du programme SPLATCHE afin de simuler les interactions entre deux populations différentes

Cette annexe présente des modifications apportées au programme SPLATCHE, afin de permettre la simulation de deux populations en interaction selon le modèle décrit dans la section 4.3. L'implémentation d'une deuxième population dans SPLATCHE est inspirée par deux études antérieures, qui ont simulé les interactions entre chasseurs-collecteurs paléolithiques et agriculteurs néolithiques en Europe (Rendine *et al.* 1986 ; Barbujani *et al.* 1995), même si notre méthodologie s'en différencie passablement sur de nombreux points que nous discutons dans la section 4.4.

Annexe 4.1 Deux matrices de dèmes superposées

Le principe de base de notre implémentation est le suivant : à chaque cellule du monde virtuel correspondent deux dèmes qui représentent chacun une des deux population (*A* ou *B*). Il est donc possible de se représenter ce monde virtuel par deux matrices de dèmes superposées (Figure 9.15). La densité des populations à l'intérieur des dèmes peut évoluer au cours du temps, de même que les dèmes peuvent échanger des migrants. Les **migrations intrapopulationnelles** (voir section 4.3.2.1) sont celles qui ont lieu à l'intérieur d'une même population (Figure 9.15), par exemple entre deux dèmes appartenant à une couche ($A \leftrightarrow A$ ou $B \leftrightarrow B$). A l'opposé, les **migrations interpopulationnelles**, qui représentent l'**hybridation** entre populations (voir section 4.3.2.2), sont celles qui ont lieu entre deux couches ($A \leftrightarrow B$, Figure 9.15).

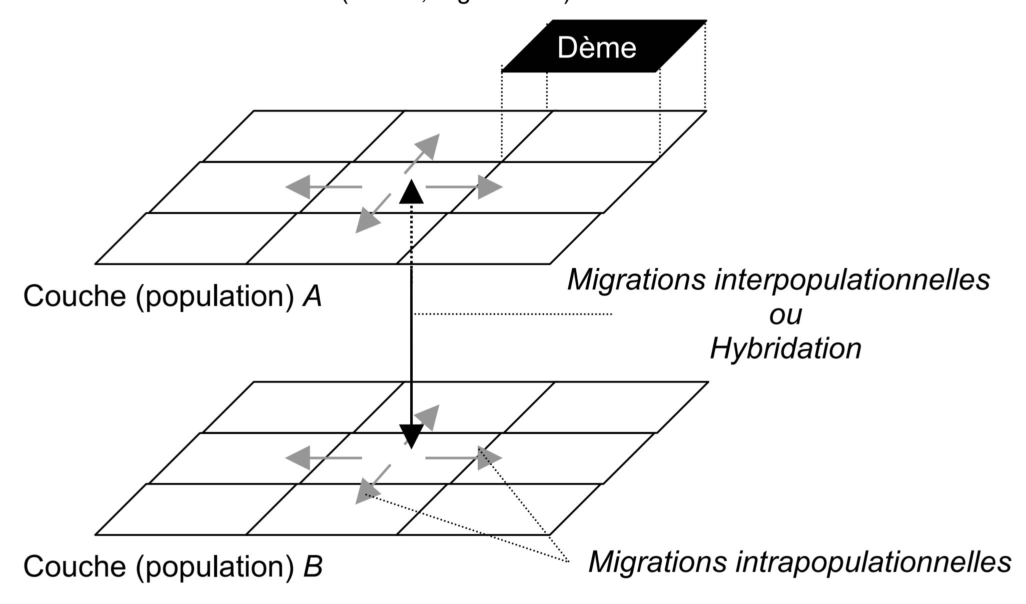


Figure 9.15 Schéma du modèle utilisé pour simuler la démographie de deux populations (*A* et *B*). Il s'agit de deux matrices de dèmes superposées, chacune représentant l'une des populations. Des migrations intrapopulationnelles sont possibles entre dèmes voisins à l'intérieur de chaque couche, et des migrations interpopulationnelles (hybridation) sont possibles entre dèmes en vis-à-vis, appartenant à des couches différentes.

Annexe 4.2 Relations ancestrales entre populations différentes

La simulation de deux populations différentes implique également l'existence de deux origines, une pour chacune d'entre-elles. Un arbre de coalescence n'ayant qu'une seule racine, il a donc fallu définir les relations ancestrales entre les populations, afin de permettre la simulation de données génétiques. Deux stratégies différentes ont été élaborées à cette fin (Figure 9.16):

- Stratégie 1 : La seconde population (*B*) est issue d'individus appartenant à la première population (*A*). Pendant la phase démographique, à un temps choisi par l'utilisateur (t = 0 dans Figure 9.16), un nombre *n* d'individus appartenant à un dème de la population *A* migre dans le dème homologue de la population *B* alors que tous les autres dèmes appartenant à la matrice représentant la population *B* sont vides. Cette stratégie illustre, par exemple, le cas du Néolithique européen (chapitre 6), puisque des chasseurs-collecteurs appartenant à des dèmes du Proche-Orient (population *A*) adoptent l'agriculture et créent la population néolithique (*B)*. Ce scénario ne requiert aucune modification particulière dans le module génétique, puisque les lignages qui arrivent dans la cellule source *B*, en remontant le temps, migrent automatiquement dans la population *A* au temps 0 et pourront ensuite coalescer dans cette population jusqu'à ce qu'il ne reste que le MRCA.
- Stratégie 2 : La population *B* est créée par des individus totalement indépendants de la population *A*, de sorte que lorsque l'on remonte le temps, les lignages peuvent se trouver soit dans le dème source de la population *A*, soit dans celui de la population *B*. L'utilisateur va donc décider d'un temps τ (en générations) pendant lequel les deux dèmes vont rester séparés, avant d'être réunis dans un seul et unique dème. Dans ce cas, l'hypothèse est donc faite que les deux populations *A* et *B* ont un ancêtre commun $\tau + t$ générations avant le présent. Cette stratégie est utilisée lors de la simulation du remplacement des Néandertaliens (chapitre 5).

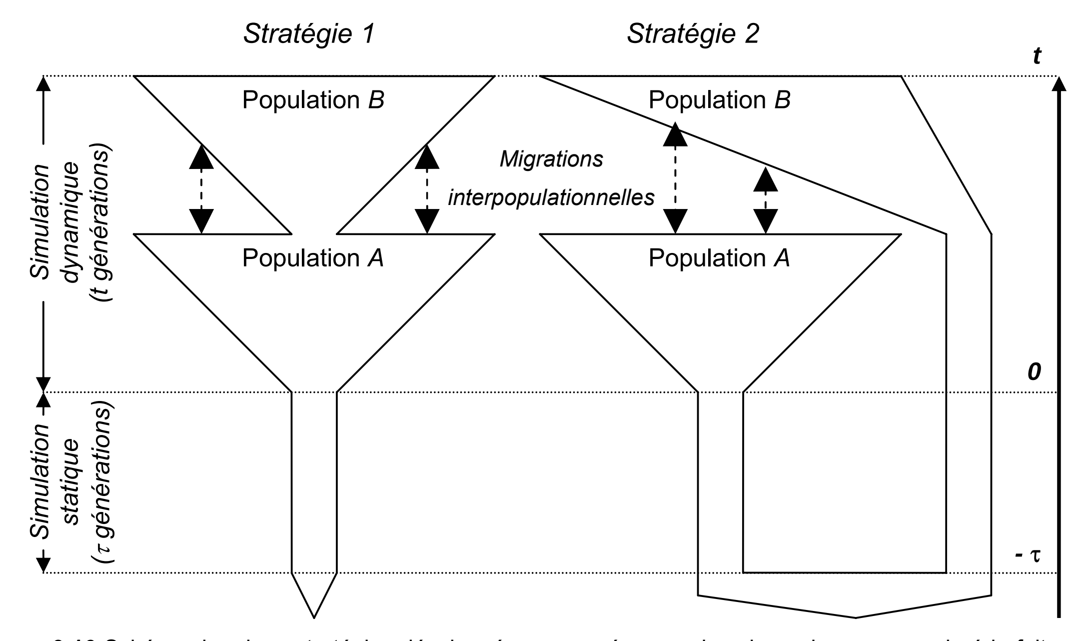


Figure 9.16 Schéma des deux stratégies développées pour créer un arbre de coalescence, malgré le fait que chaque population possède sa propre origine.

Annexe 4.3 Echantillonnage simultané dans chacune des populations

La version modifiée de SPLATCHE permet de spécifier dans quelle population un échantillon doit être tiré, soit dans la population *A*, soit dans la population *B*, soit éventuellement dans les deux à la fois comme le montre la Figure 9.17. Dans les applications présentées aux chapitres 5 et 6, l'échantillonnage est effectué uniquement dans la population survivante (Hommes modernes ou Néolithiques), mais la possibilité d'échantillonnage dans deux populations différentes peut être intéressante dans des situations pour lesquelles des données génétiques sont disponibles pour chacune des deux populations en compétition.

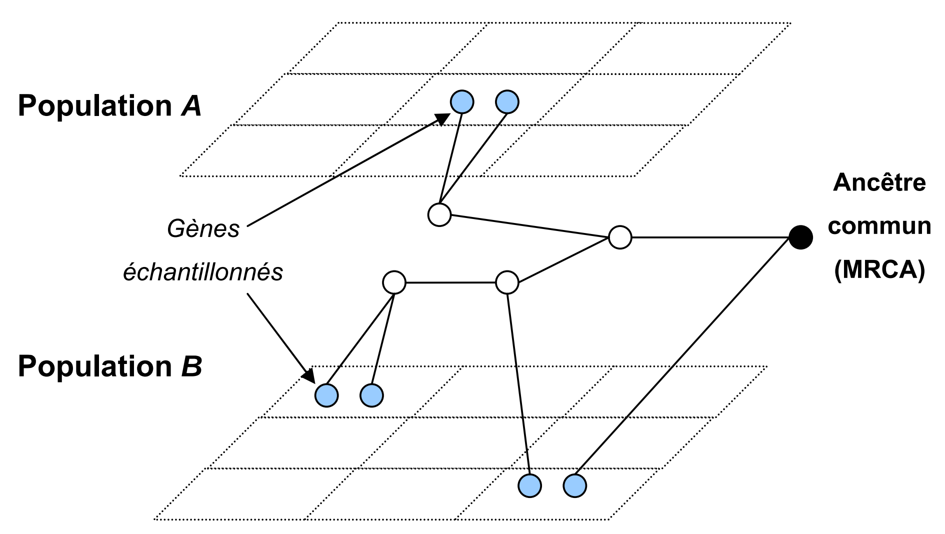


Figure 9.17 Schéma d'un arbre de coalescence pour 6 gènes, dont 2 sont tirés de la population *A* et 4 de la population *B* (2 gènes dans un dème, et 2 gènes dans un autre).

Annexe 4.4 Possibilité d'extension à *n* populations

L'implémentation d'une seconde population a été réalisée de façon à ce que, dans le futur, plus de deux populations différentes puissent évoluer simultanément dans le monde virtuel. Toutes les modifications ont été faites dans ce sens. Par exemple, à chaque cellule correspond un vecteur de dèmes, la position du dème dans le vecteur étant égale au numéro de la population (matrice) à laquelle il appartient. Il est donc aisé d'augmenter la taille du vecteur, afin de permettre la simulation d'un plus grand nombre de populations. La simulation de *n* populations, avec $n \geq 3$, à l'aide de la version modifiée de SPLATCHE nécessite cependant des modifications supplémentaires. Cela réclamerait, en effet, l'implémentation de la lecture des paramètres qui caractérisent les populations additionnelles (notamment *K* et *F*). Il faudrait surtout développer un modèle démographique qui régisse les interactions entre les *n* populations, puisque seuls des modèles incorporant une (chapitre 2) ou deux (section 4.3) populations ont été développés dans le cadre du projet "Friction".

Il est cependant clair que les modèles de compétition envisagés dans la section 4.3 peuvent être facilement étendus à un nombre arbitraire de populations.

De nombreuses autres modifications ont dû être apportées à SPLATCHE, regroupées sous une variable de compilation[1] _MULTIDEME_, afin de permettre l'évolution simultanée de deux populations différentes. Nous ne les décrirons cependant pas ici, puisque elles ne présentent aucun intérêt particulier pour le lecteur.

---

[1] Une variable de compilation est un mot clef qui spécifie au compilateur quelles sont les parties du code à utiliser.

10 Bibliographie

Les références de cette bibliographie se rapportent uniquement aux citations dans le texte. Les références déjà citées dans les publications (ou manuscrits) comprises dans ce travail (2.2.1, 3.2.1, 5.2.1 et 6.2.1 et ANNEXE 1) n'y figurent pas.

Aborgast R.-M., Magny M., Pétrequin P. (1996) Climat, cultures céréalières et densité de population au néolithique: Le cas des lacs du jura français de3500 à2500 av. J.-C. *Archäologisches Korrespondenzblatt* 26:121-143.

Adams J., Faure H. (1997) Review and atlas of palaeovegetation: preliminary land ecosystem maps of the world since Last Glacial Maximum. Oak Bridge National Laboratory.

Akey J.M., Zhang G., Zhang K., Jin L., Shriver M.D. (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 12:1805-14.

Allen J.R.M., Brandt U., Brauer A., Hubberten H.-W., Huntley B., Keller J., Kraml M., Mackensen A., Mingram J., Negendank J.F.W., Nowaczyk N.R., Oberhänsli H., Watts W.A., Wulf S., Zolitschka B. (1999) Rapid environmental changes in southern Europe during the last glacial period. *Nature* 400:740-743.

Allison A.C. (1954) Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J* 4857:290-4.

Alroy J. (2001) A multispecies overkill simulation of the end-Pleistocene megafaunal mass extinction. *Science* 292:1893-1896.

Al-Zahery N., Semino O., Benuzzi G., Magri C., Passarino G., Torroni A., Santachiara-Benerecetti A.S. (2003) Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. *Mol Phylogenet Evol* 28:458-72.

Ammerman A., Cavalli-Sforza L.L. (1971) Measuring the rate of spread of early farming in Europe. *Man* 6:674-688.

Ammerman A., Cavalli-Sforza L.L. (1984) The Neolithic transition and the genetics of populations in Europe. Princeton University Press, Princeton, New Jersey.

Anderson D.G., Gillam C. (2000) Paleoindian colonization of the Americas: implications from an examination of physiography, demography, and artifact distribution. *American Antiquity* 65:43-66.

Anderson S., Bankier A.T., Barrell B.G., de Bruijn M.H., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., Schreier P.H., Smith A.J., Staden R., Young I.G. (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290:457-65.

Aoki K. (1996) Modelling the spread of early farming and of the early upper palolithic in Europe. In: Omoto K, Tobias PV (eds) The Origins and Past of Modern Humans - Towards Reconciliation. Vol 3: Recent Advances in Human Biology. World Scientific, Kyoto, pp 206-227.

Aoki K., Shida M., Shigesada N. (1996) Travelling Wave Solutions for the Spread of Farmers into a Region Occupied by Hunter-Gatherers. *Theor Popul Biol* 50:1-17.

Appenzeller T., Clery D., Culotta E. (1998) Archeology: Transitions in Prehistory. *Science* 282:1441-1458.

Arias P. (1999) The origins of the Neolithic along the Atlantic coast of continental Europe. *Journal of World Prehistory* 13:403-464.

Aris-Brosou S., Excoffier L. (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol. Biol. Evol.* 13:494-504.

Barbujani G., Sokal R.R. (1991) Genetic population structure of Italy. II. Physical and cultural barriers to gene flow. *American Journal of Human Genetics* 48:398-411.

Barbujani G., Pilastro A. (1993) Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. *Proceedings of the National Academy of Science* 90:4670-3.

Barbujani G., Whitehead G.N., Bertorelle G., Nasidze I.S. (1994) Testing hypothesis on processes og genetic and linguistic change in the Caucasus. *Hum Biol* 66:843-864.

Barbujani G., Sokal R.R., Oden N.L. (1995) Indo-European origins: a computer-simulation test of five hypotheses. *Am J Phys Anthropol* 96:109-32.

Barbujani G., Stenico M., Excoffier L., Nigro L. (1996) Mitochondrial DNA sequence variation across linguistic and geographic boundaries in Italy. *Hum Biol* 68:201-15.

Barbujani G., Bertorelle G., Chikhi L. (1998) Evidence for Paleolithic and Neolithic gene flow in Europe. *Am J Hum Genet* 62:488-92.
Barbujani G., Chikhi L. (2000) Genetic population structure of europeans inferred from nuclear and mitochondrial DNA polymorphisms. In: Renfrew C, Boyle K (eds) Archaeogenetics:DNA and the population prehistory of Europe. Vol 1. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, pp 119-129.
Barbujani G., Bertorelle G. (2001) Genetics and the population history of Europe. *Proc Natl Acad Sci U S A* 98:22-5..
Barbujani G., Dupanloup I. (2002) DNA Variation in Europe: estimating the demographic impact of Neolithic dispersals. In: Bellwood P, Renfrew C (eds) Examining the farming/language dispersal hypothesis. McDonald Institute Monographs, Cambrige, pp 421-431.
Beaumont M., Barratt E.M., Gottelli D., Kitchener A.C., Daniels M.J., Pritchard J.K., Bruford M.W. (2001) Genetic diversity and introgression in the Scottish wildcat. *Mol Ecol* 10:319-36.
Beaumont M., Rannala B. (2004) The Bayesian revolution in genetics. *Nat Reviews | Genet* 5:251-261.
Beaumont M.A., Zhang W., Balding D.J. (2002) Approximate Bayesian Computation in Population Genetics. *Genetics* 162:2025-2035.
Begon M., Harper J.L., Townsend C.R. (1996) Ecology. Blackwell Science, Oxford.
Belledi M., Poloni E.S., Casalotti R., Conterio F., Mikerezi I., Tagliavini J., Excoffier L. (2000) Maternal and paternal lineages in Albania and the genetic structure of Indo-European populations. *Eur J Hum Genet* 8:480-6.
Bellwood P. (2001) Early Agriculturalist Population. *Annu. Rev. Anthropo.* 30:181-207.
Belovsky G.E. (1988) An optimal foraging-based model of hunter-gatherer population dynamics. *Journal of anthropological archaeology* 7:329-372.
Bentley R.A., Price T.D., Lüning J., Gronenborn D., Wahl J., Fullagar P.D. (2002) Prehistoric migration in Europe: Strontium Isotope Analysis of Early Neolithic Skeletons. *Current Anthropology* 43:799-804.
Bentley R.A., Chikhi L., Price T.D. (2003) The Neolithic transition in Europe: comparing broad scale genetic and local scale isotopic evidence. *Antiquity* 77:63-65.
Bernatchez L., Glémet H., Wilson C.C., Danzmann R.G. (1995) Introgression and fixation of Arctic char (Salvelinus alpinus) mitochondrial genome in an allopatric population of brook trout (Salvelinus fontinalis). *Canadian Journal of Fisheries and Aquatic Science* 52:179-185.
Bertranpetit J., Sala J., Calafell F., Underhill P.A., Moral P., Comas D. (1995) Human mitochondrial DNA variation and the origin of Basques. *Ann Hum Genet* 59 (Pt 1):63-81.
Binford L.R. (2001) Constructing frames of reference. An analytical method for archaeological theory building using hunter-gatherer and environmental data sets. University of California Press, Berkeley.
Biraben J.N. (1979) Essay on the evolution of numbers of mankind. *Population* 34:13-25.
Biraben J.-N. (2003) L'évolution du nombre des hommes. *Population et Sociétés* 394:1-4.
Birdsell J.B. (1957) Some population problems involving Pleistocene man. *Cold Spring Harbor Symposium on Quantitative Biology* 22:47-69.
Birdsell J.B. (1968) Some predictions for the Pleistocene based on equilibrium systems among recent hunter-gatherers. In: Lee RB, DeVore I (eds) Man the hunter. Aldine Publishing Company, Chicago, pp 229-240.
Blurton Jones N.G., Hawkes K., O'Connell J.F. (2002) Antiquity of postreproductive life: are there modern impacts on hunther-gatherer postreproductive life spans? *American Journal of Human Biology* 14:184-205.
Bocquet-Appel J.-P., Demars P.Y. (2000a) Population Kinetics in the Upper Palaeolithic in western Europe. *Journal of Archaeological Science* 27:551-570.
Bocquet-Appel J.-P., Demars P.Y. (2000b) Neanderthal contraction and modern human colonization of Europe. *Antiquity* 74:544-552.
Bosch E., Calafell F., Santos F.R., Perez-Lezaun A., Comas D., Benchemsi N., Tyler-Smith C., Bertranpetit J. (1999) Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am J Hum Genet* 65:1623-38.
Bosch E., Calafell F., Comas D., Oefner P.J., Underhill P.A., Bertranpetit J. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am J Hum Genet* 68:1019-29.
Brion M., Salas A., Gonzalez-Neira A., Lareu M.V., Carracedo A. (2003) Insights into Iberian population origins through the construction of highly informative Y-chromosome haplotypes using biallelic markers, STRs, and the MSY1 minisatellite. *Am J Phys Anthropol* 122:147-61.

Buhler S., Sanchez-Mazas A., Zanone R., Djavad N., Tiercy J.M. (2002) PCR-SSOP molecular typing of HLA-C alleles in an Iranian population. *Tissue Antigens* 59:525-30.

Calafell F., Bertranpetit J. (1993) The genetic history of the Iberian peninsula: a simulation. *Current Anthropology* 34:735-745.

Calafell F., Underhill P., Tolun A., Angelicheva D., Kalaydjieva L. (1996) From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann Hum Genet* 60 (Pt 1):35-49.

Capelli C., Redhead N., Abernethy J.K., Gratrix F., Wilson J.F., Moen T., Hervig T., Richards M., Stumpf M.P., Underhill P.A., Bradshaw P., Shaha A., Thomas M.G., Bradman N., Goldstein D.B. (2003) A y chromosome census of the british isles. *Curr Biol* 13:979-84.

Cappello N., Rendine S., Griffo R., Mameli G.E., Succa V., Vona G., Piazza A. (1996) Genetic analysis of Sardinia: I. data on 12 polymorphisms in 21 linguistic domains. *Ann Hum Genet* 60 (Pt 2):125-41.

Caramelli D., Lalueza-Fox C., Vernesi C., Lari M., Casoli A., Mallegni F., Chiarelli B., Dupanloup I., Bertranpetit J., Barbujani G., Bertorelle G. (2003) Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. *Proc Natl Acad Sci U S A* 100:6593-7.

Carvajal-Carmona L.G., Soto I.D., Pineda N., Ortiz-Barrientos D., Duque C., Ospina-Duque J., McCarthy M., Montoya P., Alvarez V.M., Bedoya G., Ruiz-Linares A. (2000) Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *Am J Hum Genet* 67:1287-95.

Casalotti R., Simoni L., Belledi M., G. B. (1999) Y-chromosome polymorphisms and the origins of the European gene pool. *Proc R Soc Lond B Biol Sci* 266:1959-1965.

Cavalli-Sforza L. (1996) The spread of agriculture and nomadic pastoralism: Insights from genetics, linguistics and archealogy. In: Harris DR (ed) The Origins and Spread of Agriculture and Pastoralism in Eurasia. UCL Press, London, pp 51-70.

Cavalli-Sforza L.L., Hewlett B. (1982) Exploration and mating range in African Pygmies. *Ann Hum Genet* 46:257-70.

Cavalli-Sforza L.L., King M.C. (1986) Detecting linkage for genetically heterogeneous diseases and detecting heterogeneity with linkage data. *Am J Hum Genet* 38:599-616.

Cavalli-Sforza L.L., Piazza A. (1993) Human genomic diversity in Europe: a summary of recent research and prospects for the future. *Eur J Hum Genet* 1:3-18.

Cavalli-Sforza L.L., Menozzi P., Piazza A. (1994) The History and Geography of Human Genes. In. Princeton University Press, Princeton, New Jersey, pp 145-154.

Cavalli-Sforza L.L., Minch E. (1997) Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 61:247-54.

Cavalli-Sforza L.L., Feldman M.W. (2003) The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 33 Suppl:266-75.

Chakraborty R. (1975) Estimation of race admixture- A new method. *American Journal of Physical Anthropology* 42:507-511.

Chikhi L., Destro-Bisol G., Bertorelle G., Pascali V., Barbujani G. (1998) Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. *Proc Natl Acad Sci U S A* 95:9053-8.

Chikhi L., Bruford M.W., Beaumont M.A. (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* 158:1347-62.

Chikhi L. (2002) Admixture and the demic diffusion model in Europe. In: Bellwood P, Renfrew C (eds) Examining the farming/language dispersal hypothesis. McDonald Institute Monographs, Cambrigs, pp 435-447.

Chikhi L., Nichols R.A., Barbujani G., Beaumont M.A. (2002) Y genetic data support the Neolithic demic diffusion model. *Proc Natl Acad Sci U S A* 99:11008-13.

Clark J.G.D. (1965) Radiocarbon dating and thee expansion of farming culture from the Near East over Europe. *Proc. prehist. Soc.* 31:58-73.

Coale A.J. (1974) The history of the human population. *Scientific American* 231:40-51.

Cockerham C. (1969) Variance of gene frequenceies. *Evolution* 23:72-84.

Cockerham C. (1973) Analyse of gene frequencies. *Genetics* 74:679-700.

Comas D., Calafell F., Mateu E., Perez-Lezaun A., Bertranpetit J. (1996) Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Mol Biol Evol* 13:1067-77.

Comas D., Calafell F., Mateu E., Perez-Lezaun A., Bosch E., Bertranpetit J. (1997) Mitochondrial DNA variation and the origin of the Europeans. *Hum Genet* 99:443-9.

Comas D., Calafell F., Bendukidze N., Fananas L., Bertranpetit J. (2000) Georgian and kurd mtDNA sequence analysis shows a lack of correlation between languages and female genetic lineages. *Am J Phys Anthropol* 112:5-16.
Conard J.N., Bolus M. (2003) Radiocarbon dating the appearance of modern humans and timing of cultural innovations in Europe: new results and new challenges. *Journal of Human Evolution* 44:331-371.
Corte-Real H.B., Macaulay V.A., Richards M.B., Hariti G., Issad M.S., Cambon-Thomsen A., Papiha S., Bertranpetit J., Sykes B.C. (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* 60:331-50.
Crawford M.H. (1998) The origins of Native Americans. Cambridge University Press, Cambridge.
Currat M. (1999) Etude de la variabilité allozymiques des Valaisans et des Walsers. Diplôme de Biologie, Université de Genève, Genève.
Currat M., Trabuchet G., Rees D., Perrin P., Harding R.M., Clegg J.B., Langaney A., Excoffier L. (2002) Molecular Analysis of the beta-Globin Gene Cluster in the Niokholo Mandenka Population Reveals a Recent Origin of the betaS Senegal Mutation. *Am J Hum Genet* 70:207-223.
Dard P., Schreiber Y., Excoffier L., Sanchez-Mazas A., Shi-Isaac X., Epelbouin A., Langaney A., Jeannet M. (1992) [Polymorphism of HLA class I loci HLA-A, -B, -C, in the Mandenka population from eastern Senegal]. *C R Acad Sci III* 314:573-8.
Dard P., Lefranc M.-P., Osipova L., Sanchez-Maza A. (2001) DNA sequence variability of IGHG3 alleles associated to the main G3m haplotypes in human populations. *European Journal of Human Genetics* 9:765-777.
de Knijff P., Kayser M., Caglia A., Corach D., Fretwell N., Gehrig C., Graziosi G., et al. (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. *Int J Legal Med* 110:134-49.
de Menocal P.B. (2001) Cultural responses to climate change during the late Holocene. *Science* 292:667-73.
Delghandi M., Utsi E., Krauss S. (1998) Saami mitochondrial DNA reveals deep maternal lineage clusters. *Hum Hered* 48:108-14.
Di Rienzo A., Wilson A.C. (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc Natl Acad Sci U S A* 88:1597-1601.
Diamond J., Bellwood P. (2003) Farmers and their languages: the first expansions. *Science* 300:597-603.
Djindjian F., Koslowski J., Otte M. (1999) Le Paléolithique supérieur en Europe. Armand Colin, Paris.
Donnelly P., Tavaré S. (1995) Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29:401-421.
Duarte C., Mauricio J., Pettitt P.B., Souto P., Trinkaus E., van der Plicht H., Zilhao J. (1999) The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. *Proc Natl Acad Sci U S A* 96:7604-9.
Dugoujon J.-M., Hazout S., Loirat F., Mourrieras B., Crouau-Roy B., Sanchez-Mazas A. (2004) GM haplotype diversity of 82 populations over the world suggests a centrifugal model of human migrations. *Am J Phys Anthropol* 123.
Dunn F.L. (1968) Epidemiological Factors: Health and Disease in Hunther-Gatherer. In: Lee RB, DeVore I (eds) Man the hunter. Aldine Publishing Company, Chicago, pp 221-227.
Dupanloup de Ceuninck I. (1999) Evaluation et synthèse des contributions de la linguistique et de la génétique à l'étude de la différenciation des populations humaines pendant la préhistoire récente. Thèse, Université de Genève, Genève.
Dupanloup I., Pereira L., Bertorelle G., Calafell F., Prata M.J., Amorim A., Barbujani G. (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *J Mol Evol* 57:85-97.
Edmonds C.A., Lillie A.S., Cavalli-Sforza L.L. (2004) Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci U S A* 101:975-9.
ESRI (1998) ARCVIEW 3.1. Environmental Systems Research Institute.
Eswaran V. (2002) A diffusion wave out of Africa - the mechanism of the modern human revolution. *in prep*:1-53.
Ewens W.J. (1990) Population Genetics Theory - The Past and the Future. In: Lessar S (ed) Mathematical and Statistical developments of Evolutionary Theory. Kluwer Academic Publishers, Dordrecht, pp 177-227.

Excoffier L., Pellegrini B., Sanchez-Mazas A., Simon C., Langaney A. (1987) Genetics and history of Sub-Saharan Africa. *Yearbook of Physical Anthropology* 30:151-194.
Excoffier L. (1988) Polymorphisme de l'ADN mitochondrial et histoire du peuplement humain. Thèse, Université de Genève, Genève.
Excoffier L., Harding R.M., Sokal R.R., Pellegrini B., Sanchez-Mazas A. (1991) Spatial Differentiation of RH and GM Haplotype Frequencies in Sub-Saharan Africa and Its Relation to Linguistic Affinities. *Human Biology* 63:273-307.
Excoffier L., Smouse P., Quattro J. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479-491.
Excoffier L. (1997) Ce que nous dit la généalogie des gènes. *La Recherche* N° 302:82-90.
Excoffier L., Schneider S. (1999) Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc Natl Acad Sci U S A* 96:10597-602.
Excoffier L., Novembre J., Schneider S. (2000) SIMCOAL: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J. Heredity* 91:506-510.
Excoffier L. (2002) Human demographic history: refining the recent African origin model. *Curr Opin Genet Dev* 12:675-82.
Excoffier L. (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol Ecol* 13:853-64.
Fiedel S.J. (1992) Prehistory of the Americas. Cambridge University Press, Cambridge.
Fiedel S.J., Anthony D.W. (2003) Deerslayers, pathfinders, and iceman. In: Rockman M, Steele J (eds) Colonization of unfamiliar landscapes: The archaeology of adaptation. Routledge, London, pp 104-168.
Fix A.G. (1996) Gene frequency clines in Europe: demic diffusion or natural selection? *J. Roy. anthrop. Inst.* 2:625-643.
Fix A.G. (1997) Gene frequency clines produced by kin-structured founder effects. *Hum Biol* 69:663-73.
Flores J.C. (1998) A mathematical model for Neanderthal extinction. *J theor Biol* 191:295-298.
Forster P., Rohl A., Lunnemann P., Brinkmann C., Zerjal T., Tyler-Smith C., Brinkmann B. (2000) A short tandem repeat-based phylogeny for the human Y chromosome. *Am J Hum Genet* 67:182-96.
Forster P., Cali F., Röhl A., Metspalu E., D'Anna R., Mirisola M., De Leo G., Flugy A., Salerno A., Ayala G., Kouvatsi A., Villems R., Romano V. (2002) Continental and subcontinental distributions of mtDNA control region types. *Int J Legal Med* 116:99-108.
Francalacci P., Bertranpetit J., Calafell F., Underhill P.A. (1996) Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am J Phys Anthropol* 100:443-60.
Fu Y.-X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. *Genetics* 147:915-925.
Gallay A. (1994) A propos de travaux récents sur la Néolithisation de l'Europe de l'ouest. *L'Anthropologie* 98:576-588.
Gallay A. (2004) A propos du livre de Karoline Mazurié de Keroualin : Genèse et diffusion de l'agriculture en Europe: agriculteurs, chasseurs, pasteurs. (Paris: Errance 2003). *in press*.
Gilmour J.S.L., Gregor J.W. (1939) Demes: a suggested new terminology. *Nature* 144:333.
Gkiasta M., Russell T., Shennan S., Steele J. (2003) Neolithic transition in Europe: the radiocarbon record revisited. *Antiquity* 77:45-62.
Gribchenko Y.N., Kurenkova E.I. (1999) Pleistocene environments and the dispersal of Paleolithic groupes in eastern Europe. *Anthropologie* 37:79-87.
Grimaud-Hervé D., Serre F., Bahain J.-J., Nespoulet R. (2001) Histoire d'ancêtres: la grande aventure de la préhistoire. Artcom', Paris.
Gronenborg D. (1999) A variation on a basic theme: the transition to farming in southern central Europe. *Journal of World Prehistory* 13:123-210.
Hagelberg E. (2003) Recombination or mutation rate heterogeneity? Implications for Mitochondrial Eve. *Trends Genet* 19:84-90.
Haldane J.B.S. (1949) Disease and evolution. *La Ricerca Scientifica (Suppl.)* 19:68-76.
Hamilton G., Currat M., Ray N., Heckel G., Beaumont M., Excoffier L. (2004) Bayesian estimation of recent migration rates after a spatial expansion. *submitted*.

Hammer M.F., Karafet T., Rasanayagam A., Wood E.T., Altheide T.K., Jenkins T., Griffiths R.C., Templeton A.R., Zegura S.L. (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol* 15:427-41.
Hammer M.F., Karafet T.M., Redd A.J., Jarjanazi H., Santachiara-Benerecetti S., Soodyall H., Zegura S.L. (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* 18:1189-203.
Hammer M.F., Blackmer F., Garrigan D., Nachman M.W., Wilder J.A. (2003) Human population structure and its effects on sampling Y chromosome sequence variation. *Genetics* 164:1495-509.
Handt O., Meyer S., von Haeseler A. (1998) Compilation of human mtDNA control region sequences. *Nucleic Acids Research* 26:126-12.
Harpending H., Sherry S.T., Rogers A.R., Stoneking M. (1993) The genetic structure of ancient human populations. *Current Anthropology* 34:483-496.
Harpending H. (2001) Book reviews: Archaeogenetics: DNA and the population prehistory of Europe. *Am J Phys Anthropol* 116:177-179.
Harpending H.C. (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Hum Biol* 66:591-600.
Harpending H.C., Batzer M.A., Gurven M., Jorde L.B., Rogers A.R., Sherry S.T. (1998) Genetic traces of ancient demography. *Proc Natl Acad Sci U S A* 95:1961-7..
Harris D.R. (1996) The origins and spread of agriculture and pastoralism in Eurasia. University College London Press, London.
Hartl D.L., Clark A.G. (1997) Principles of Population Genetics. Sinauer Associates, Inc, Sunderland, Massachusetts.
Hassan F.A. (1979) Demography and archaeology. *Annual Review of Anthropology* 8:137-160.
Hassan F.A. (1981) The peopling of the World. In: Demographic archaeology. Academic Press, New York, pp 193-208.
Hazelwood L., Steele J. (2003) Colonizing new landscape. In: Rockman M, Steele J (eds) Colonization of unfamiliar landscapes: The archaeology of adaptation. Routledge, London, pp 203-221.
Helgason A., Hickey E., Goodacre S., Bosnes V., Stefansson K., Ward R., Sykes B. (2001) mtDna and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *Am J Hum Genet* 68:723-37.
Helgason A., Hrafnkelsson B., Gulcher J.R., Ward R., Stefansson K. (2003) A Populationwide Coalescent Analysis of Icelandic Matrilineal and Patrilineal Genealogies: Evidence for a faster Evolutionary Rate of mtDNA Lineages than Y Chromosomes. *Am J Hum Genet* 72:00-00.
Hewitt G. (2000) The genetic legacy of the quartenary ice ages. *Nature* 405:907-913.
Hewitt G.M. (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* 58:247-276.
Hewitt G.M. (2001) Speciation, hybrid zones and phylogeography - or seeing genes in space and time. *Molecular Ecology* 10:537-549.
Hewlett B., Van de Koppel J.M.H., Cavalli-Sforza L. (1982) Exploration ranges of Aka pygmies of the Central African Republic. *Man* 17:418-430.
Hill E.W., Jobling M.A., Bradley D. (2000a) Y-chromosome variation and Irish origins. In: Renfrew C, Boyle K (eds) Archaeogenetics:DNA and the population prehistory of Europe. Vol 1. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, pp 81-88.
Hill E.W., Jobling M.A., Bradley D.G. (2000b) Y-chromosome variation and Irish origins. *Nature* 404:351-2.
Horai S., Hayasaka K. (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. *Am J Hum Genet* 46:828-42.
Housley R.A., Gamble C.S., Street M., Pettit P. (1997) Radiocarbon evidence for the lateglacial human recolonisation of northern Europe. *Proc. Prehist. Soc.* 63:25-54.
Hublin J.-J. (1988) Le peuplement paléolithique de l'Europe: un point de vue paléobiogéographique. Paper presented at Colloque international de Nemours. Nemours.
Hublin J.-J. (2002) Demographic crashes in Pleistocene Europe and Neanderthal evolution. Paper presented at Human origins & disease. Cold Spring Harbor, New York.
Hudson R.R. (1990) Gene genealogies and the coalescent process. Vol 7. Oxford University Press, oxford.

Huntley B. (1988) Glacial and holocene vegetation history - 20 ky to present. Europe. In: Huntley B, Webb T (eds) Vegetation history. Vol 7: Handbook of Vegetation Science. Kluwer Academic Publishers, pp 341-383.
Hurles M.E., Veitia R., Arroyo E., Armenteros M., Bertranpetit J., Perez-Lezaun A., Bosch E., Shlumukova M., Cambon-Thomsen A., McElreavey K., Lopez De Munain A., Rohl A., Wilson I.J., Singh L., Pandya A., Santos F.R., Tyler-Smith C., Jobling M.A. (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am J Hum Genet* 65:1437-48.
Hurles M.E., Jobling M.A. (2001) Haploid chromosomes in molecular ecology: lessons from the human Y. *Mol Ecol* 10:1599-613.
Hurles M.E., Nicholson J., Bosch E., Renfrew C., Sykes B.C., Jobling M.A. (2002) Y chromosomal evidence for the origins of oceanic-speaking peoples. *Genetics* 160:289-303.
Hurles M.E., Maund E., Nicholson J., Bosch E., Renfrew C., Sykes B.C., Jobling M.A. (2003) Native american y chromosomes in polynesia: the genetic impact of the polynesian slave trade. *Am J Hum Genet* 72:1282-7.
Hyden B. (1990) *J. Anthrop. Archaeol.* 9:31.
Jackes M., Lubell D., Meiklejohn C. (1997) Healthy but mortal: human biology and the first farmers of western Europe. *Antiquity* 71:639-658.
Jeunesse C. (1998) La néolithisation de l'Europe occidentale (VIIe-Ve millénaires av. J.-C.): nouvelles perspectives. In: Cupillard C, Richard A (eds) Les derniers chasseurs-cueilleurs du massif jurassien et de ses marges. Centre Jurassien du patrimoine, Lons-le-Saunier.
Jobling M.A., Hurles M.E., Tyler-Smith C. (2004) Human evolutionary genetics: origins, peoples and disease. Garland Science, New York.
Jorde L.B., Bamshad M.J., Watkins W.S., Zenger R., Fraley A.E., Krakowiak P.A., Carpenter K.D., Soodyall H., Jenkins T., Rogers A.R. (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *Am J Hum Genet* 57:523-38.
Kaessmann H., Zollner S., Gustafsson A.C., Wiebe V., Laan M., Lundeberg J., Uhlen M., Paabo S. (2002) Extensive linkage disequilibrium in small human populations in eurasia. *Am J Hum Genet* 70:673-85.
Kalmar T., Bachrati C.Z., Gyorgypal Z., Downes C.S., Rasko I. (2003) Mitochondrial lineages in Hungarian-speaking populations of the Carpathian basin. Vol. 2003. Genbank.
Kaplan N., Hudson R.R., Iizuka M. (1991) The coalescent process in models with selection, recombination and geographic subdivision. *Genet Res* 57:83-91.
Kayser M., Krawczak M., Excoffier L., Dieltjes P., Corach D., Pascali V., Gehrig C., Bernini L.F., Jespersen J., Bakker E., Roewer L., de Knijff P. (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am J Hum Genet* 68:990-1018..
Kimura M. (1953) "Stepping-stone" model of population. *Annual Report of National Institute of Genetics* 3:62-63.
King J.P., Kimmel M., Chakraborty R. (2000) A power analysis of microsatellite-based statistics for inferring past population growth. *Mol Biol Evol* 17:1859-68.
Kingman J.F.C. (1982a) The coalescent. *Stoch. Proc. Appl.* 13:235-248.
Kingman J.F.C. (1982b) On the genealogy of large populations. *J. Appl. Proba.* 19A:27-43.
Kittles R.A., Bergen A.W., Urbanek M., Virkkunen M., Linnoila M., Goldman D., Long J.C. (1999) Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. *Am J Phys Anthropol* 108:381-99.
Klein R.G. (2003) Paleoanthropology. Whither the Neanderthals? *Science* 299:1525-7.
Klopfstein S. (in prep.) Travail de diplôme, Université de Berne, Berne.
Kozlowski J., Otte M. (2000) The formation of the Aurignacian. *Journal of Anthropological Research* 56:513-524.
Krings M., Stone A., Schmitz R.W., Krainitzki H., Stoneking M., Paabo S. (1997) Neandertal DNA sequences and the origin of modern humans. *Cell* 90:19-30.
Krings M., Geisert H., Schmitz R.W., Krainitzki H., Paabo S. (1999) DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. *Proc Natl Acad Sci U S A* 96:5581-5.
Krings M., Capelli C., Tschentscher F., Geisert H., Meyer S., von Haeseler A., Grossschmidt K., Possnert G., Paunovic M., Paabo S. (2000) A view of Neandertal genetic diversity. *Nat Genet* 26:144-6.
Laan M., Paabo S. (1997) Demographic history and linkage disequilibrium in human populations. *Nat Genet* 17:435-8.

Lahr M.M., Foley R.A. (1998) Toward a Theory of Modern Human Origins: Geography, Demography, and Diversity in Recent Human Evolution. *Yearbook of Physical Anthropology* 41:137-176.
Landers J. (1992) Reconstructing ancient populations. In: Jones S, Martin R, Pilbeam D (eds) The Cambridge Encyclopedia of Human Evolution. Cambrige University Press, London, pp 402-405.
Langaney A. (1988) Les Hommes, passé, présent, conditionnel. Armand Colin, Paris.
Langaney A., Hubert Van Blyenburgh N., Nadot R. (1990) L'histoire génétique des mille derniers siècles et ses mécanismes: une revue. *Bull. et Mém. de la Soc. d'Anthrop. de Paris* 1:43-56.
Langaney A., Roessli D., Hubert Van Blyenburgh N., Dard P. (1992) Do most human populations descend from phylogenetic trees. *Human Evolution* 2:47-61.
Larruga J.M., Diez F., Pinto F.M., Flores C., Gonzalez A.M. (2001) Mitochondrial DNA characterisation of European isolates: the Maragatos from Spain. *Eur J Hum Genet* 9:708-16.
Laval G., Excoffier L. (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*.
Lee R.B., DeVore I. (1968a) Problems in the study of hunters and gatherers. In: Lee RB, DeVore I (eds) Man the hunter. Aldine Publishing Company, Chicago, pp 4-12.
Lee R.B., DeVore I. (1968b) Man the hunter. Aldine Publishing Company, Chicago.
Lev-Yadun S., Gopher A., Abbo S. (2000) Archaeology. The cradle of agriculture. *Science* 288:1602-3.
Lewontin R.C., Kojima K. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution* 14:450-472.
Lewontin R.C. (1988) On measures of gametic disequilibrium. *Genetics* 120:849-52.
Livingstone F.B. (1989) Simulation of the diffusion of the B-globin variants in the old world. *Human Biology* 61 (3):297-309.
Lordkipanidze D. (1999) The settlements of mountainous regions: a view from the Caucasus. *Anthropologie* 37:71-78.
Lotka A.J. (1932) The growth of mixed populations: two species competing for a common food supply. *Journal of the Washington academy of Sciences* 22:461-469.
Lubbock J. (1865) Prehistoric times, as illustrated by ancient remains and the manners and customs of modern savages. Williams and Norgate, London.
Lundstrom R., Tavare S., Ward R.H. (1992) Modeling the evolution of the human mitochondrial genome. *Math Biosci* 112:319-35.
Malaspina P., Cruciani F., Ciminelli B.M., Terrenato L., Santolamazza P., Alonso A., Banyko J., Brdicka R., Garcia O., Gaudiano C., Guanti G., Kidd K.K., Lavinha J., Avila M., Mandich P., Moral P., Qamar R., Mehdi S.Q., Ragusa A., Stefanescu G., Caraghin M., Tyler-Smith C., Scozzari R., Novelletto A. (1998) Network analyses of Y-chromosomal types in Europe, northern Africa, and western Asia reveal specific patterns of geographic distribution. *Am J Hum Genet* 63:847-60.
Malécot G. (1948) Les Mathématiques de l'Hérédité. Masson, Paris.
Malécot G. (1955) The decrease of relationship with distance. *Cold Spring Harbor Symp. Quant. Biol.* 20:52-53.
Malyarchuk B.A., Derenko M.V. (2001) Mitochondrial DNA variability in Russians and Ukrainians: implication to the origin of the Eastern Slavs. *Ann Hum Genet* 65:63-78.
Marjoram P., Donnelly P. (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* 136:673-683.
Mazurié de Keroualin K. (2001) La première néolithisation de l'Europe: une réévaluation des modalités du peuplement. Thèse, Université de Genève, Genève.
Mazurié de Keroualin K. (2003) Genèse et diffusion de l'agriculture en Europe : agriculteurs, chasseurs, pasteurs. Errance, Paris.
Mellars P. (1998) The fate of the Neanderthals. *Nature* 395:539-40.
Mellars P.A. (1992) Archaeology and the population-dispersal hypothesis of modern human origins in Europe. *Philos Trans R Soc Lond B Biol Sci* 337:225-34.
Menozzi P., Piazza A., Cavalli-Sforza L. (1978) Synthetic maps of human gene frequencies in Europeans. *Science* 201:786-92.
Mogentale-Profizi N., Chollet L., Stevanovitch A., Dubut V., Poggi C., Pradie M.P., Spadoni J.L., Gilles A., Beraud-Colomb E. (2001) Mitochondrial DNA sequence diversity in two groups of Italian Veneto speakers from Veneto. *Ann Hum Genet* 65:153-66.
Morton N.E. (1977) Isolation by distance in human populations. *Ann Hum Genet* 40:361-5.

Morton N.E. (1982) Estimation of demographic parameters from isolation by distance. *Hum. Hered.* 32:37-41.
Mosimann J.F., Martin P.S. (1975) Simulating overkill by paleoindians. *American Scientist* 63:304-313.
Mountain J.L., Hebert J.M., Bhattacharyya S., Underhill P.A., Ottolenghi C., Gadgil M., Cavalli-Sforza L.L. (1995) Demographic history of India and mtDNA-sequence diversity. *Am J Hum Genet* 56:979-92.
Mourant A.E., Kopec A.C., Domaniewska-Sobczak K. (1976) The distribution of the human blood groups and others polymorphisms, sd edition. Oxford University Press, London.
Nasidze I., Stoneking M. (2001) Mitochondrial DNA variation and language replacements in the Caucasus. *Proc R Soc Lond B Biol Sci* 268:1197-206.
Neuenschwander S. (in prep.) The simulation of the colonization history of European bullhead (*Cottus gobio L.*) across the Rhine-Rhône watershed in Switzerland. Thèse de doctorat, Université de Berne, Berne.
Nichols R. (2001) Gene trees and species trees are not the same. *Trends Ecol Evol* 16:358-364.
Nichols R.A., Hewitt G.M. (1994) The genetic consequences of long distance dispersal during colonization. *Heredity* 72:312-317.
Nordborg M. (1998) On the probability of Neanderthal ancestry. *Am J Hum Genet* 63:1237-40.
Nordborg M. (2001) Coalescent theory. In: Balding D, Bishop C, Cannings C (eds) Handbook of Statistical Genetics. John Wiley & Sons Ltd, New York, pp 179-212.
Notohara M. (1990) The coalescent and the genealogical process in geographically structured population. *J Math Biol* 29:59-75.
Oota H., Settheetham-Ishida W., Tiwawech D., Ishida T., Stoneking M. (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nat Genet* 29:20-1.
Oota H., Kitano T., Jin F., Yuasa I., Wang L., Ueda S., Saitou N., Stoneking M. (2002) Extreme mtDNA homogeneity in continental Asian populations. *Am J Phys Anthropol* 118:146-53.
Oppenheimer S., Richards M. (2001) Fast trains, slow boats, and the ancestry of the Polynesian islanders. *Sci Prog* 84:157-81.
Orekhov V., Poltoraus A., Zhivotovsky L.A., Spitsyn V., Ivanov P., Yankovsky N. (1999) Mitochondrial DNA sequence diversity in Russians. *FEBS Lett* 445:197-201.
Otte M. (2000) The history of european populations as seen by archaeology. In: Renfrew C, Boyle K (eds) Archaeogenetics:DNA and the population prehistory of Europe. Vol 1. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, pp 41-44.
Ovchinnikov I.V., Götherström A., Romanova G.P., Kharitonov V.M., Liden K., Goodwin W. (2000) Molecular analysis of Neanderthal DNA from the nothern Caucasus. *Nature* 404:490-493.
Parson W., Parsons T.J., Scheithauer R., Holland M.M. (1998) Population data for 101 Austrian caucasian mitochondrial DNA d-loop sequences: Application of mtDNA sequence analysis to forensic case. *Int J Legal Med* 111:124-132.
Pennington R. (2001) Hunter-gatherer demography. In: Panter-Brick C, Layton RH, Rowley-Conwy P (eds) Hunter-gatherers: an interdisciplinary perspective. Cambridge University Press, pp 170-204.
Pereira L., Dupanloup I., Rosser Z.H., Jobling M.A., Barbujani G. (2001) Y-chromosome mismatch distributions in Europe. *Mol Biol Evol* 18:1259-71.
Piazza A., Rendine S., Minch E., Menozzi P., Mountain J., Cavalli-Sforza L.L. (1995) Genetics and the origin of European languages. *Proc Natl Acad Sci U S A* 92:5836-40.
Piercy R., Sullivan K.M., Benson N., Gill P. (1993) The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int J Legal Med* 106:85-90.
Pinhasi R., Foley R.A., Lahr M.M. (2000) Spatial and temporal patterns in the Mesolithic-Neolithic archaeological record of Europe. In: Renfrew C, Boyle K (eds) Archaeogenetics:DNA and the population prehistory of Europe. Vol 1. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, pp 45-56.
Poloni E.S. (1991) Le peuplement de la Chine: hypothèses linguistiques, archéologiques et génétiques. Diplôme de Biologie, Université de Genève, Genève.
Poloni E.S., Excoffier L., Mountain J.L., Langaney A., Cavalli-Sforza L.L. (1995) Nuclear DNA polymorphism in a Mandenka population from Senegal: comparison with eight other human populations. *Ann Hum Genet* 59:43-61.
Poloni E.S., Semino O., Passarino G., Santachiara-Benerecetti A.S., Dupanloup I., Langaney A., Excoffier L. (1997) Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet* 61:1015-35.

Poloni E.S. (1999) Polymorphisme de l'ADN et histoire du peuplement humain: apport de l'étude des marqueurs RFLP. Thèse, Université de Genève, Genève.
Pritchard J.K., Seielstad M.T., Perez-Lezaun A., Feldman M.W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 16:1791-8.
Pult I., Sajantila A., Simanainen J., Georgiev O., Schaffner W., Paabo S. (1994) Mitochondrial DNA sequences from Switzerland reveal striking homogeneity of European populations. *Biol Chem Hoppe Seyler* 375:837-40.
Quintana-Murci L., Semino O., Minch E., Passarimo G., Brega A., Santachiara-Benerecetti A.S. (1999) Further characteristics of proto-European y chromosomes. *Eur J Hum Genet* 7:603-8.
Quintana-Murci L., Veitia R., Fellous M., Semino O., Poloni E.S. (2003) Genetic structure of Mediterranean populations revealed by Y-chromosome haplotype analysis. *Am J Phys Anthropol* 121:157-71.
Ray N., Adams J.M. (2001) A GIS-based vegetation map of the world at the Last Glacial Maximum (25,000-15,000 BP). *Internet Archaeology* 11.
Ray N., Adams J. (2002) Vegetation maps of Europe at four key time intervals, part of the article "Les climats de l'Europe". *National Geographic France (in press)*.
Ray N. (2003) Modélisation de la démographie des populations humaines préhistoriques à l'aide de données environnementales et génétiques. Thèse, Université de Genève, Genève.
Ray N., Currat M., Excoffier L. (2003) Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol* 20:76-86.
Ray N., M. C., Excoffier L. (2004) Simulating realistic genetic diversity to find the origin of a population expansion. *in prep*.
Reich D.E., Schaffner S.F., Daly M.J., McVean G., Mullikin J.C., Higgins J.M., Richter D.J., Lander E.S., Altshuler D. (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 32:135-42.
Relethford J.H. (2001) Absence of regional affinities of Neandertal DNA with living humans does not reject multiregional evolution. *Am J Phys Anthropol* 115:95-8.
Rendine S., Piazza A., Cavalli-Sforza L. (1986) Simulation and separation by principal components of multiple demic expansions in Europe. *Am. Nat.* 128:681-706.
Renfrew C. (1989) Archaeology and Language: The Puzzle of Indo-European Origins. Penguin Books, London.
Renfrew C. (2000) Archaeogenetics: Towards a Population Prehistory of Europe. In: Renfrew C, Boyle K (eds) Archaeogenetics:DNA and the population prehistory of Europe. Vol 1. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, pp 3-12.
Renquin J., Sanchez-Mazas A., Halle L., Rivalland S., Jaeger G., Mbayo K., Bianchi F., Kaplan C. (2001) HLA class II polymorphism in Aka Pygmies and bantu Congolese and reassessment of HLA-DRB1 African diversity. *Tissue Antigens* 58:211-222.
Reynolds J., Weir B.S., Cockerham C.C. (1983) Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105:767-779.
Richards M., Corte-Real H., Forster P., Macaulay V., Wilkinson-Herbots H., Demaine A., Papiha S., Hedges R., Bandelt H.J., Sykes B. (1996) Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 59:185-203.
Richards M., Macaulay V., Hickey E., Vega E., Sykes B., Guida V., Rengo C., et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* 67:1251-76..
Richards M., Macaulay V., Torroni A., Bandelt H.J. (2002) In search of geographical patterns in European mitochondrial DNA. *Am J Hum Genet* 71:1168-74.
Richards M. (2003) The Neolithic invasion of Europe. *Annu. Rev. Anthropol.* 32:135-162.
Richards M., Rengo C., Cruciani F., Gratrix F., Wilson J.F., Scozzari R., Macaulay V., Torroni A. (2003) Extensive female-mediated gene flow from sub-saharan Africa into near eastern arab populations. *Am J Hum Genet* 72:1058-64.
Richards M.B., Macaulay V.A., Bandelt H.J., Sykes B.C. (1998) Phylogeography of mitochondrial DNA in western Europe. *Ann Hum Genet* 62 (Pt 3):241-60.
Roebroeks W. (2001) Hominid behaviour and the earliest occupation of Europe: an exploration. *Journal of Human Evolution* 41:437-461.
Roebroeks W. (2003) Landscape learning and the earliest peopling of Europe. In: Rockman M, Steele J (eds) Colonization of unfamiliar landscapes: The archaeology of adaptation. Routledge, London, pp 99-115.

Roewer L., Kayser M., de Knijff P., Anslinger K., Betz A., Caglia A., Corach D., Furedi S., Henke L., Hidding M., Kargel H.J., Lessig R., Nagy M., Pascali V.L., Parson W., Rolf B., Schmitt C., Szibor R., Teifel-Greding J., Krawczak M. (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Sci Int* 114:31-43.
Rogers A.R., Harpending H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* 9:552-69.
Rogers A.R., Jorde L.B. (1996) Ascertainment bias in estimates of average heterozygosity. *Am J Hum Genet* 58:1033-41.
Rosser Z.H., Zerjal T., Hurles M.E., Adojaan M., Alavantic D., Amorim A., Amos W., et al. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* 67:1526-43.
Rousselet F., Mangin P. (1998) Mitochondrial DNA polymorphisms: a study of 50 French Caucasian individuals and application to forensic casework. *Int J Legal Med* 111:292-8.
Rousset F. (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* 142:1357-1362.
Roychoudhury A.K., Nei M. (1988) Human Polymorphic Genes World Distribution. Vol 1, New york - Oxford.
Sagart L., Blench R., Sanchez-Mazas A. (2004) The Peopling of East Asia: putting together Archaeology, Linguistics and Genetics. RoutledgeCurzon, London.
Sajantila A., Lahermo P., Anttinen T., Lukka M., Sistonen P., Savontaus M.L., Aula P., Beckman L., Tranebjaerg L., Gedde-Dahl T., Issel-Tarver L., DiRienzo A., Paabo S. (1995) Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res* 5:42-52..
Sajantila A., Paabo S. (1995) Language replacement in Scandinavia. *Nat Genet* 11:359-60.
Sajantila A., Salem A.H., Savolainen P., Bauer K., Gierig C., Paabo S. (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc Natl Acad Sci U S A* 93:12035-9.
Sanchez-Mazas A. (1990) Polymorphisme des systèmes immunologiques Rhésus, GM et HLA et histoire du peuplement humain. Thèse, Université de Genève, Genève.
Sanchez-Mazas A., Bütler-Brunner E., Excoffier L., Ghanem N., Ben-Salem M., Breguet G., Dard P., Pellegrini B., Tikkanen J., Lefranc G., Langaney A., Bütler R. (1994) New data for AG haplotype frequencies in Caucasoid populations and the selective neutrality of the AG polymorphism. *Hum Biol* 66:27-48.
Sanchez-Mazas A. (2000) The Berbers of North Africa: Genetic Relationships according to HLA and other polymorphisms. In: Arnaiz-Villena A (ed) Prehistoric Iberia: Genetics, Anthropology, and Linguistics. Kluwer Academic/Plenum, New York, pp 65-77.
Sanchez-Mazas A. (2001a) Les origines de l'homme, au coeur de ses gènes. *Pour la science* 289:84-91.
Sanchez-Mazas A. (2001b) African diversity from the HLA point of view: influence of genetic drift, geography, linguistics, and natural selection. *Hum Immunol* 62:937-48.
Schillaci M.A., Froehlich J.W. (2001) Nonhuman primate hybridization and the taxonomic status of Neanderthals. *Am J Phys Anthropol* 115:157-66.
Schmitz R.W., Serre D., Bonani G., Feine S., Hillgruber F., Krainitzki H., Paabo S., Smith F.H. (2002) The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc Natl Acad Sci U S A* 99:13342-7.
Schneider S., Roessli D., Excoffier L. (2000) Arlequin: a software for population genetics data analysis. User manual release 2.000, Geneva.
Scholz M., Bachmann L., Nicholson G.J., Bachmann J., Giddings I., Rüschoff B., Czarnetzki A., Push C.M. (2000) Genomic differentiation of Neanderthals and Anatomically modern man allows a fossil-DNA-based classification of morphologically indistinguishable hominid bones. *Am J Hum Genet* 66:1927-1932.
Schwartz J., Tattersall I. (1996) Significance of Some Previously Unrecognized Apomorphies in the Nasal Region of Homo neanderthalensis. *Proc Natl Acad Sci U S A* 93:10852-10854.
Seielstad M.T., Minch E., Cavalli-Sforza L.L. (1998) Genetic evidence for a higher female migration rate in humans. *Nat Genet* 20:278-80.
Semino O., Passarino G., Brega A., Fellous M., Santachiara-Benerecetti A.S. (1996) A view of the neolithic demic diffusion in Europe through two Y chromosome-specific markers. *Am J Hum Genet* 59:964-8.
Semino O., Passarino G., Oefner P.J., Lin A.A., Arbuzova S., Beckman L.E., De Benedictis G., Francalacci P., Kouvatsi A., Limborska S., Marcikiae M., Mika A., Mika B., Primorac D.,

Santachiara-Benerecetti A.S., Cavalli-Sforza L.L., Underhill P.A. (2000a) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* 290:1155-9.
Semino O., Passarino G., Quintana-Murci L., Liu A., Beres J., Czeizel A., Santachiara-Benerecetti A.S. (2000b) MtDNA and Y chromosome polymorphisms in Hungary: inferences from the palaeolithic, neolithic and Uralic influences on the modern Hungarian gene pool. *Eur J Hum Genet* 8:339-46.
Serre D., Langaney A., Chech M., Teschler-Nicola M., Paunovic M., Mennecier P., Hofreiter M., Possnert G.G., Paabo S. (2004) No Evidence of Neandertal mtDNA Contribution to Early Modern Humans. *PLoS Biol* 2:E57.
Sgaramella-Zonta L., Cavalli-Sforza L. (1973) A methode for the detection of a demic cline. In: Morton NE (ed) Genetic structure of population: Population Genetics Monograph 3. University of Hawaii Press, Honolulu, HI.
Shastry B.S. (2002) SNP alleles in human disease and evolution. *J Hum Genet* 47:561-6.
Shaw K.L. (2002) Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: What mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proc Natl Acad Sci U S A* 99:16122-16127.
Shen P., Wang F., Underhill P.A., Franco C., Yang W.H., Roxas A., Sung R., Lin A.A., Hyman R.W., Vollrath D., Davis R.W., Cavalli-Sforza L.L., Oefner P.J. (2000) Population genetic implications from sequence variation in four Y chromosome genes. *Proc Natl Acad Sci U S A* 97:7354-9.
Shen P., Buchholz M., Sung R., Roxas A., Franco C., Yang W.H., Jagadeesan R., Davis K., Oefner P.J. (2002) Population genetic implications from DNA polymorphism in random human genomic sequences. *Hum Mutat* 20:209-17.
Sherratt A. (1997) Climatic cycles and behavioural revolutions: the emergence of modern humans and the beginning of farming. *Antiquity* 71:271-287.
Simoni L., Gueresi P., Pettener D., Barbujani G. (1999) Patterns of gene flow inferred from genetic distances in the Mediterranean region. *Hum Biol* 71:399-415..
Simoni L., Calafell F., Pettener D., Bertranpetit J., Barbujani G. (2000) Geographic patterns of mtDNA diversity in Europe. *Am J Hum Genet* 66:262-78.
Skaletsky H., Kuroda-Kawaguchi T., Minx P.J., Cordum H.S., Hillier L., Brown L.G., Repping S., et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423:825-37.
Slatkin M. (1991) Inbreeding coefficients and coalescence times. *Genet. Res. Camb.* 58:167-175.
Slatkin M., Hudson R.R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555-62.
Slatkin M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457-462.
Sokal R.R., Menozzi P. (1982) Spatial Autocorrelations of HLA Frequencies in Europe Support Demic Diffustion of Early Farmers. *American Naturalist* 119:1-17.
Sokal R.R. (1988) Genetic, geographic, and linguistic distances in Europe. *Proceedings of the National Academy of Science* 85:1722-1726.
Sokal R.R., Oden N.L., Thomson B.A. (1988) Genetic changes across language boundaries in Europe. *American Journal of Physical Anthropology* 76:337-61.
Sokal R.R. (1991a) Ancient movement patterns determine modern genetic variances in Europe. *Human Biology* 63:589-606.
Sokal R.R. (1991b) The Continental Population Structure of Europe. *Annual Review of Anthropology* 20:119-140.
Sokal R.R., Jacquez G.M. (1991) Testing inferences about microevolutionary processes by means of spatial autocorrelation analysis. *Evolution* 45:152-168.
Sokal R.R., Oden N.L., Wilson C. (1991) Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature* 351:143-5.
Sokal R.R., Jacquez G.M., Oden N.L., DiGiovanni D., Falsetti A.B., McGee E., Thomson B.A. (1993) Genetic relationships of European populations reflect their ethnohistorical affinities. *Am J Phys Anthropol* 91:55-70.
Sokal R.R., Oden N.L., Walker J., Di Giovanni D., Thomson B.A. (1996) Historical population movements in Europe influence genetic relationships in modern samples. *Hum Biol* 68:873-98.
Spielmann K.A., Eder J.F. (1994) Hunters and Farmers: then and now. *Annu. Rev. Anthropol.* 23:303-323.

Steele J., Adams J.M., Sluckin T. (1998) Modeling Paleoindian dispersals. *World Archeology* 30:286-305.
Stefan M., Stefanescu G., Gavrila L., Terrenato L., Jobling M.A., Malaspina P., Novelletto A. (2001) Y chromosome analysis reveals a sharp genetic boundary in the Carpathian region. *Eur J Hum Genet* 9:27-33.
Stringer C. (1989) The Origin of Early Moden Humans: a Comparison of thr European and non-European Evidence. In: Mellars P, Stringer C (eds) The Human Revolution: Biological perspectives in the Origins of Modern Humans. Princeton University Press, Princeton, pp 233-244.
Stringer C., Grun R. (1991) Palaeoanthropology. Time for the last Neanderthals. *Nature* 351:701-2.
Stringer C., Davies W. (2001) Archaeology. Those elusive Neanderthals. *Nature* 413:791-2.
Stringer C.B., Andrews P. (1988) Genetic and fossil evidence for the origin of modern humans. *Science* 239:1263-8.
Sykes B. (1999) The molecular genetics of European ancestry. *Philos Trans R Soc Lond B Biol Sci* 354:131-8; discussion 138-9..
Taberlet P., Fumagalli L., Wust-Saucy A.G., Cosson J.F. (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Mol Ecol* 7:453-64.
Tajima F. (1989a) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.
Tajima F. (1989b) The effect of change in population size on DNA polymorphism. *Genetics* 123:597-601.
Tattersall I., Schwartz J.H. (1999) Hominids and hybrids: the place of Neanderthals in human evolution. *Proc Natl Acad Sci U S A* 96:7117-7119.
Thorpe I.J. (1999) The Origins of Agriculture in Europe. Routledge, New York.
Tills D., Kopec A.C. (1983) The distribution of the human blood groups and other polymorphisms. Oxford University Press, Oxford.
Tolan-Smith C. (2003) The social context of landscape learning and the lateglacial-early postglacial recolonization of the brithish isles. In: Rockman M, Steele J (eds) Colonization of unfamiliar landscapes: The archaeology of adaptation. Routledge, London, pp 116-129.
Torroni A., Bandelt H.J., D'Urbano L., Lahermo P., Moral P., Sellitto D., Rengo C., Forster P., Savontaus M.L., Bonne-Tamir B., Scozzari R. (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet* 62:1137-52..
Torroni A., Bandelt H.J., Macaulay V., Richards M., Cruciani F., Rengo C., Martinez-Cabrera V., et al. (2001) A Signal, from Human mtDNA, of Postglacial Recolonization in Europe. *Am J Hum Genet* 69:844-52..
Tremblay M., Vezina H. (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *American Journal of Human Genetics* 66:651-8.
Tsoularis A., Wallace J. (2002) Analysis of logistic growth models. *Math Biosci* 179:21-55.
Underhill P.A., Shen P., Lin A.A., Jin L., Passarino G., Yang W.H., Kauffman E., Bonne-Tamir B., Bertranpetit J., Francalacci P., Ibrahim M., Jenkins T., Kidd J.R., Mehdi S.Q., Seielstad M.T., Wells R.S., Piazza A., Davis R.W., Feldman M.W., Cavalli-Sforza L.L., Oefner P.J. (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* 26:358-61.
Underhill P.A., Passarino G., Lin A.A., Shen P., Mirazon Lahr M., Foley R.A., Oefner P.J., Cavalli-Sforza L.L. (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. hum. Genet.* 65:43-62.
Van Andel T.H. (2000) Where received wisdom fails: the Mid-Palaeolithic and early Neolithic climates. In: Renfrew C, Boyle K (eds) Archaeogenetics:DNA and the population prehistory of Europe. Vol 1. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, pp 31-39.
Verhulst P.F. (1838) Notice sur la loi que la population suit dans son accroissement. *Curr. Math. Phys.* 10:113.
Vignal A., Milan D., SanCristobal M., Eggen A. (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genet Sel Evol* 34:275-305.
Volterra V. (1926) Variations and fluctuations of the numbers of individuals in animal species living together (Reprinted in 1931). In: Chapman RN (ed) Animal Ecology. Mc Graw Hill, New York.
Wahl L.M., Gerrish P.J., Saika-Voivod I. (2002) Evaluating the impact of population bottlenecks in experimental evolution. *Genetics* 162:961-71.

Wakeley J. (1999) Nonequilibrium migration in human history. *Genetics* 153:1863-71.
Wakeley J. (2000) The effects of subdivision on the genetic divergence of populations and species. *Evolution* 54:1092-1101.
Wakeley J. (2001) The coalescent in an island model of population subdivision with variation among demes. *Theor Popul Biol* 59:133-44.
Wakeley J., Aliacar N. (2001) Gene genealogies in a metapopulation. *Genetics* 159:893-905.
Wall J.D. (2000) Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* 154:1271-9.
Watson E., Bauer K., Aman R., Weiss G., von Haeseler A., Paabo S. (1996) mtDNA sequence diversity in Africa. *Am J Hum Genet* 59:437-44.
Weidenreich F. (1946) Apes, giants and man. University of Chicago Press, Chicago.
Weiss K.M. (1984) On the number of members of the Genus Homo who have ever lived, and some evolutionary implications. *Human Biology* 56:637-649.
Whitfield L.S., Sulston J.E., Goodfellow P.N. (1995) Sequence variation of the human Y chromosome. *Nature* 378:379-80.
Whittle A. (1996) Europe in the Neolithic: the creation of new worlds. Cambridge University Press, Cambridge, UK.
Willis K.J., Whittaker R.J. (2000) Perspectives: paleoecology. The refugial debate. *Science* 287:1406-7.
Winterhalder B., Baillargeon W., Cappelletto F. (1988) The population ecology of hunter-gathers and their prey. *Journal of Anthropological Archaeology* 7:289-328.
Wolpoff M. (1996) Human Evolution. McGraw-Hill, New York.
Wolpoff M.H. (1989) Multiregional evolution: the fossil alternative to Eden. In: Mellars P, Stringer C (eds) The Human Revolution: Biological perspectives in the Origins of Modern Humans. Princeton University Press, Princeton, pp 62-108.
Wright S. (1943) Isolation by distance. *Genetics* 28:114-138.
Young D.A., Bettinger R.L. (1995) Simulating the global human expansion in the late pleistocene. *Journal of Archaeological Science* 22:89-92.
Zane L., Bargelloni L., Patarnello T. (2002) Strategies for microsatellite isolation: a review. *Mol Ecol* 11:1-16.
Zhivotovsky L.A., Rosenberg N.A., Feldman M.W. (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am J Hum Genet* 72:1171-86.
Zvelebil K.V. (1989) On the transition to farming in Europe, or what was spreading with the Neolithic : a reply to Ammerman. *Antiquity* 63:379-382.
Zvelebil M. (1986) Review of Ammerman & Cavalli-Sforza (1984). *Journal of Archaeological Science* 13:93-95.
Zvelebil M., Zvelebil K.V. (1988) Agricultural transition and Indo-European dispersals. *Antiquity* 62:574-583.
Zvelebil M. (2000) The social context of the Agricultural transition in Europe. In: Renfrew C, Boyle K (eds) Archaeogenetics:DNA and the population prehistory of Europe. Vol 1. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, pp 57-79.